PEER INFLUENCE, INFORMATION QUALITY AND PREDICTIVE POWER OF STOCK MICROBLOGS

by

Chong Keat Oh

A dissertation submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Business Administration

David Eccles School of Business

The University of Utah

May 2013

STATEMENT OF DISSERTATION APPROVAL

The dissertation of **Chong Keat Oh**

has been approved by the following supervisory committee members:

Olivia Sheng	, Chair	**3/8/2013** Date Approved
Xiao Fang	, Member	**3/8/2013** Date Approved
Rohit Aggarwal	, Member	**3/8/2013** Date Approved
Vandana Ramachandran	, Member	**3/8/2013** Date Approved
Daniel Zeng	, Member	**3/19/2013** Date Approved

and by **William Hesterly** , Associate Dean of the **David Eccles School of Business**

and by Donna M. White, Interim Dean of The Graduate School.

ABSTRACT

Due to the popularity of Web 2.0 and Social Media in the last decade, the percolation of user generated content (UGC) has rapidly increased. In the financial realm, this results in the emergence of virtual investing communities (VIC) to the investing public. There is an on-going debate among scholars and practitioners on whether such UGC contain valuable investing information or mainly noise.

I investigate two major studies in my dissertation. First I examine the relationship between peer influence and information quality in the context of individual characteristics in stock microblogging. Surprisingly, I discover that the set of individual characteristics that relate to peer influence is not synonymous with those that relate to high information quality. In relating to information quality, influentials who are frequently mentioned by peers due to their name value are likely to possess higher information quality while those who are better at diffusing information via retweets are likely to associate with lower information quality. Second I propose a study to explore predictability of stock microblog dimensions and features over stock price directional movements using data mining classification techniques. I find that author-ticker-day dimension produces the highest predictive accuracy inferring that this dimension is able to capture both relevant author and ticker information as compared to author-day and ticker-day. In addition to these two studies, I also explore two topics: network structure of co-tweeted tickers and sentiment annotation via crowdsourcing. I do this in order to

understand and uncover new features as well as new outcome indicators with the objective of improving predictive accuracy of the classification or saliency of the explanatory models. My dissertation work extends the frontier in understanding the relationship between financial UGC, specifically stock microblogging with relevant phenomena as well as predictive outcomes.

To my loving family

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express my deepest appreciation to my advisor and committee chair, Dr. Olivia Sheng, whose continuous guidance and mentoring these last 5 years played a pivotal role in making this dissertation a possibility and her immense knowledge in data mining and research that spur the thrust of my work. I am also grateful to my advisor, Dr. Rohit Aggarwal, who planted the seed of rigor and inquisitiveness in me, and introduced me to the world of microblogging. I am also grateful to my other committee members, namely Dr. Vandana Ramachandran, Dr. Xiao Fang and Dr. Daniel Zeng, for their friendship and valuable mentoring.

I am grateful for the continuous support of my lovely wife, Windia and children, Shannon, Shane and Shayna, who are the meaning of my life. They have been my pillar of light in times of darkness and an endearing motivation to successfully complete this dissertation. Similarly I acknowledge the support of my mother and brothers, Andy and Pacer, in Malaysia. I dedicate this work to my wonderful family.

I acknowledge the support of my peers in the Information Systems program at University of Utah, namely Han-Fen Hu, Lionel Li and Iljoo Kim, who as comrades-in-arms provided the much needed companionship throughout the program.

Chong Oh

CHAPTER 1

INTRODUCTION

Motivation

Due to the increasing popularity of Web 2.0 (Ullrich et al., 2008) and social media in the last decade, the percolation of user generated content (UGC) has increased rapidly. Initially UGC were limited to low interaction, asynchronous, text-based mediums such as emails, message board postings, online bulletin boards and online product reviews. Of late however, UGC have permeated the web via highly interactive, fast pace, high volume and entertaining media such as weblogs, microblogs, videoblogs and blogs thus becoming increasingly prevalent in every thread of the social and business circles. Popularity of social sharing sites such as Facebook, Twitter and Youtube are great examples. In fact more people are on the internet than ever and social media usage is growing. Pew Research Center reported that 83% adults are using the internet, out of which 71% accessing video sharing sites, 68% accessing social-media sites, and 14% on Twitter (USA Today, 2012). As a consequence from the emergence of such UGC channels, I heed the call for a greater understanding of their implications towards research and practice.

The challenge in making sense of these nascent UGC channels results in a highly active playground of research inquiry and industry explorations. The implications are far-

reaching, permeating across various disciplines such as marketing, finance, economics, psychology, sociology, computer science and information systems. There is growing excitement and passion to investigate the relationships, value, antecedents and phenomena that are contained within these UGC. For example, marketing scholars seek better indicators of product success through investigating information diffusion, influence and word of mouth (WOM) (Aral, 2011; Cha et al., 2010; Chevalier & Mayzlin, 2006; Ghose & Han, 2011; Jansen et al., 2009; Susarla et al., 2011). Finance and economic scholars examine stock message boards and stock market indicators such as return, volatility and trade volume (Antweiler & Frank, 2004; Das & Chen, 2004). IS scholars investigate relationships, sentiment, social network and identify experts within such communities (Bollen et al., 2010; Chen et al., 2009; Forman et al., 2008; Ghose & Ipeirotis, 2011; Kwak et al., 2010; Oh & Sheng, 2011). Aggarwal et al. (2012a) investigate the influence of blog electronic WOM on venture financing. They discovered that eWOM of popular bloggers helps ventures in getting higher funding and valuations and that the impact of negative eWOM is more than positive eWOM. In fact Aggarwal et al. (2012b) concluded that negative blogs may act as a catalyst that can exponentially increase readership. In the area of politics, scholars examine microblogs for sentiment preference towards certain political candidate or policy (Poor et al., 2011). Such keen focus of academic scholarships is a case in point to the need to further understand the intricacies of UGC.

In the financial realm, the Web 2.0 hype introduces virtual investing communities (VIC) (Chen et al., 2009) to the investing public. VICs such as Yahoo Finance's stock message board, Motley Fools and Stocktwits are publishing relevant and valuable UGC

data such as investment recommendations and proprietary analysis. UGC in these channels enriches investors' ability to make better investment decisions by allowing investors to monitor the thought process and decision-making processes of others (Sprenger & Welpe, 2010). Thus, it is imperative for researchers and practitioners to understand how individuals in virtual communities interact with one another and how these behaviors relate to future predictive outcomes.

In this dissertation I examine stock microblogging, a nascent VIC and extension of the popular microblogging Twitter, in investigating relationships between features of stock microblogging, with corresponding author and market characteristics with interesting phenomena such as peer influence and information quality. In parallel, I also examine the predictive power of these features in relation to investor sentiment and future financial outcomes. This stream of research is intriguing because it reveals many characteristics that affect phenomena that were empirically challenging to investigate in the past. One example is the study of investor behaviors with financial theories involving sentiment and other psychological biases. Without a doubt, the high volume and rapid streaming of messages pose an exciting, rich and interesting source of information that is very relevant and interesting. Hence it is puzzling to note why these data is still unpopular with academia. Specifically, I select Stocktwits to collect the data for this dissertation. Stocktwits is a social, stock microblogging service that was established in October 2008. It is a variant platform of Twitter that aggregates only stock-related stock microblog postings. This service allows users to monitor the activities of traders and investors, contribute to the conversation and build reputation as savvy market wizards (TechCrunch, 2010). It is a marketplace for investing ideas that allow amateurs to interact

freely with professionals. It currently has more than 150,000 subscribers who spend an average of 32 minutes and contribute 20,000 messages per day (Stocktwits, 2012). Not surprisingly Stocktwits was awarded Time magazine's top 50 websites for 2010 (Time, 2010). The following are the two main topics in this dissertation:

1. Examine individual characteristics, peer influence and information quality in stock microblogging.
2. Analyze predictive power of stock tweet features with future stock price movements.

These two topics explore groups of features of stock microblogs to understand their dynamics in financial markets and social networks through different lens. The first topic investigates author characteristics in stock microblogging community that affect peer influence and information quality through the lens of motivation for virtual community participation (Dholakia et al., 2004). The second topic addresses a model construction and evaluation approach in extracting dimensions and features of stock tweets, author and ticker information in relating to future financial outcomes.

In essence, this dissertation offers the following research questions:

1. Do influential individuals have high information quality?
2. Which author characteristics relate to peer influence and which relate to information quality?
3. Which dimensions of stock tweets (author-day, ticker-day, author-ticker-day) have high predictive power?
4. Which features of stock tweets have higher predictive power than others?

Overview

An overview of the two topics is as follows:

The first topic seeks to explain the relationships between individual characteristics of stock microblogging authors and two pertinent outcomes: peer influence and quality of investing information, which I term intrinsic information quality as defined by Wang & Strong (1996). This is based on the virtual community participation framework as discussed in Dholakia et al. (2004). The authors applied theories of group norms (Postmes et al., 2000) and social identity (Tajfel, 1978) and concluded that people are motivated to participate in virtual communities by self-referent and group referent values. These consist of purposive values, self-discovery, maintaining interpersonal connectivity and social enhancements. This describes the motivations for stock microblogging, primarily pertains to investors who are focused and motivated to seek and share investing information, to monitor others activities, to interact and to build his/her reputation in the community (Techcrunch, 2010). This chapter concludes that individual characteristics which are salient towards peer influence are not synonymous with those salient towards information quality.

The second topic proposes a design path to improve accuracy and usability in evaluating the predictive power of stock microblog dimensions and features over stock price directional movements using data mining classification techniques. The analysis is based on 360,000 microblog postings collected over 6 months from a popular stock microblogging channel pertaining to 4570 stock tickers and 8935 distinct authors. This study reveals that the author-ticker-day dimension produced the best accuracy as compared to author-day and ticker-day dimensions. This is due to the ability of author-

ticker-day dimension to capture both relevant author and ticker information. Subsequently I find that a classification model with market/ticker, author characteristics and sentiment/opinion feature set shows high predictive accuracy signifying that these feature sets have strong predictive power over future stock price movements. I present this approach as a baseline for other predictive studies in the microblogging domain, such as those from marketing, politics, health and social studies. In so doing I heed the call of Agarwal and Lucas (2005) in explaining the transformational impact of a nascent IT artifact, stock microblogging, in connecting to reference disciplines. Specifically, I provide evidence for the model of irrational investor sentiment, recommend a supplementary investigative approach using user-generated content (UGC) for investors and a framework that may contribute to the monetization schemes for Virtual Investing Communities (VIC) for managers.

In addition to the two main topics, I explore other topics to better understand microblog features and their impact. The first explores network structure of co-tweeted tickers in extracting relationships among tickers that are co-cited in the same microblogs. Such relationships among tickers may lead to inferences from investors in relation to groups of tickers that move together in the stock market dynamics. The second explores manual annotation of sentiment via crowd sourcing in examining approaches and best practices in crowd sourcing involving nonexperts. Manual labeling is important as it provides the base for automatic labeling of investor sentiment or opinion. Without sentiment it is impossible to determine outcomes such as predictive accuracy of stock price movement or information quality.

Contributions

This dissertation has two main contributions. First, I improve the peer influence model by extending peer influence research with information quality. Information quality is a critical factor of peer influence research since the primary motivation for online participation is to seek and share information (Wasko & Faraj, 2005). However, even though scholars have examined individual characteristics (e.g., Aral, 2011; Iyengar et al., 2010) and network characteristics (e.g., Goldenberg et al., 2009) in relation to peer influence, they have largely ignored the scrutiny of information quality of exchanged information, probably due to empirical constraints. Thus the inclusion of information quality may sharpen the understanding of peer influence since intuitively those deemed with higher information quality should be more influential. Although the context of this study is in relation to stock market performance and economic outcomes, I assert that the presented models are generalizable to other research disciplines related to the microblogging domain such as those of movie, product and company tweets. Along with this new knowledge, I explore various nascent peer influence measures related to information quality. Essentially I extend the current popular peer influence microblogging measures of retweets, mentions and replies (Cha et al., 2010) by incorporating the dimensions of reciprocity and normalization which provide a shaper representation of peer influence. Reciprocity accounts for outdegree of the same peer influence type (e.g., residual of indegree and outdegree retweets) while normalization accounts for the effort of sending total tweets (i.e., residual/total tweets). In addition, I use the individual characteristic measures of self-disclosed demographics and trading

preferences, recognition and engaging effort to determine which traits are more salient to information quality versus those that are more salient towards peer influence.

My second contribution involves establishing a model construction approach to evaluate microblogs in improving accuracy and usability based on dimensions and features. This framework is applicable to other domains or industries that are active in the microblogging channel such as politics, marketing and health. Specifically related to IS researcher, I study and explain the transformational impact of a nascent IT artifact (Agarwal & Lucas, 2005), stock microblogging, to the research and practitioner communities. I also provide additional evidence to the scholars of behavioral finance in supporting the tenets of irrational investor behavior in explaining stock market movements (De Long et al., 1990; Tetlock, 2007). Furthermore, I help to clarify the influence of sentiment for UGC scholars. In addition, I propose an additional investing mechanism, using investor sentiment, to aid private as well as institutional investors in investing decision making. Finally, I identify predictive value embedded in investor sentiment that may aid in monetization schemes to managers of VICs such as Yahoo Finance or StockTwits.

On the practitioner front, both studies present new knowledge to platform managers and individual investors. Aral (2011) noted that a vast majority of data available to firms and governmental organizations are observational, making the improved understanding of causal peer influence estimation in such data critical. Better monitoring of peer influence and information quality allude to monetization opportunities for VIC managers and outline mechanisms for investors seeking influential or high quality peers. These different types of peer interactions such as retweet, reply, mention

and following unveil different aspects of peer influence (e.g. retweet relates to influence of content while mention relates to influence of the source) in providing a richer set of information for better financial decision making.

CHAPTER 2

ABOUT STOCK MICROBLOGGING AND SENTIMENT ANALYSIS

About Stock Microblogging

This chapter provides background information on stock microblogging and sentiment analysis which are the foundation for all the topics in this dissertation.

Why People Participate in Virtual Communities

In investigating VIC, the question of participant motivation has always been the underlying theme. Understanding this motivation helps to better comprehend the relationships between individual characteristics, peer influence and information quality. Social influence research provides a good framework to guide us in this study. One example is Dholakia et al. (2004) who examined virtual community participation from seven internet venues. They applied theories of group norms (Postmes et al., 2000) and social identity (Tajfel, 1978) and concluded that people are motivated by self-referent and group referent values. Self-referent values consist of 1) *purposive value* "derived from accomplishing some predetermined instrumental purpose" (i.e., giving and receiving information) and 2) *self-discovery* (i.e., understanding of one's preferences, tastes, and deepening salient aspects of one's self). Group referent values, on the other hand, consist

of 3) *maintaining interpersonal connectivity* (i.e., social support, friendship and intimacy) and 4) *social enhancements* (i.e., acceptance and approval from others and enhancement of social status or reputation within the community). Figure 1 illustrates this framework.

Stock Microblogging

Stock microblogging is a variant of the popular microblogging channel Twitter. It features a stream of on-going conversations (tweets) posted by investors, continuously highlighting the current trending investing topic. Each tweet is limited to 140 characters and the succinct content of these tweets cover opinions on various investment instruments, analysis of stocks, predictions, news, links to articles, technical charts and other stock market related information (Techcrunch, 2010). In addition, tweets may contain questions, seeking confirmation on investing decisions or even rumors. Stock microblogging presents an opportunity to explore and extend existing peer influence relationships to the financial domain, a research stream that is interesting and relevant due to the following reasons.

First, stock microblogging represents live conversations on stocks (Sprenger & Welpe, 2010) alluding to the notion of conversation discourse (Zhang & Swanson, 2010). Features of stock microblogging such as succinct, real-time, rapid, and high volume (Sprenger & Welpe, 2010) distinctively differentiate this channel from other VIC such as stock message boards. Lack of face to face interactions is substituted by high volume and real-time tweets alluding to the same feeling of co-presence (Golfman, 1967, p.17), which is defined as “persons sense that they are close enough to be perceived in whatever they are doing, including their experiencing of others, and close enough to be perceived

in this sensing of being perceived". Co-presence relates to accountability which influences the information quality of tweets shared among peers. In addition, these tweets parallel stock trading activities of online investors in tandem with market fluctuations. Such parallelism may attenuate the saliency of peer influence contained in the exchange of tweets among members of the stock microblogging channel. Also market dynamics changes constantly so new information is sought after frequently. Due to conversational content in stock microblogging the influence effect should be more salient as compared to non-conversational content (Leskovec et al., 2007). Clearly this channel offers a rich dataset that is different from other VIC where streaming of messages is less rapid and of lower volume.

Second, stock microblogging offers a rich dataset of social interactions among online investors. Prior studies state that empirical evidence regarding real world influence is limited (Watts & Dodds, 2007). However, with stock microblogging, due to Web 2.0, online social interactions are now ubiquitous and can be mapped (Katona et al., 2011). Although adoption decisions are not observed from influence of shared information, influence can be inferred by social features of retweets, mentions, and replies. These features are strong motivations for users to publish quality investing information in order to build and maintain relationships in the community (Sprenger & Welpe, 2010). In fact stock microblogging has been acknowledged for its accountability and transparency where each author's reputation is continuously scrutinized on a daily basis (Zeledon, 2009). Although salient, these features have yet to be fully understood by the research and practitioner communities.

What Investors Do in the Stock Microblogging Community

Seek and Share Investing Information (Purposive Value)

The primary objective of participating in stock microblogging is to seek and share investing information, which accumulates in informational and instrumental values (coined by Dholakia et al. (2004) as purposive value) and are key drivers for participation in virtual communities (Dholakia et al., 2004). Individuals share information by directing his/her tweets to all (public tweet) or to a specific individual (reply tweet). A public tweet is the default while a reply tweet begins with the '@' character followed by the username of the recipient (e.g. @chongoh) at the beginning of the message. Although both public and reply tweets are visible to all followers, replies are focused on the individual that the tweet is addressing, analogous to addressing an individual in a group conversation (Twitter, 2012).

As with Twitter, stock microblogging uses the same framework of followership where a follower is an individual who follows another and a following is one who is being followed. Having a larger follower network might imply a larger audience that should lead to a higher level of recognition and peer influence. Nevertheless, this assumption was rebutted in a recent study (Cha et al, 2010). In addition to public and reply tweets, an individual may also send URLs in the content of tweets that link to richer information such as a news article, blog, video or chart. Furthermore, individuals seek information primarily through reciprocation. Reciprocity is a sense of mutual indebtedness and reinforces trust (Chai et al., 2012) by reciprocating the benefits received from others and ensuring ongoing supportive exchanges (Shumaker & Brownell, 1984;

Wasko & Faraj, 2005). Essentially, individuals are motivated to broadcast high quality tweets in order to benefit from future reciprocity.

Build Reputation (Social Enhancement)

Another important objective of participating in stock microblogging is to build one's reputation as savvy market investors (Techcrunch, 2010), as per the value of social enhancement (Dholakia et al., 2004). Individuals desire to increase followers and to acquire firm recognition (i.e., the 'suggested badge') from Stocktwits. Online participation strongly motivates the cultivation of reputation (Wasko & Faraj, 2005). The rewards of having a good reputation include promoting one's investment services, monetizing tweet streams and in fulfilling one's psychological needs of being accepted as experts (Flanagin & Metzger, 2001). Intuitively, one may infer that information quality from influential experts should be high due to the motivation to build a sustainable level of reputation (Bolton et al., 2004; Resnick, 2000).

Disclose Demographics and Trading Preferences (Maintaining Interpersonal Connectivity)

One way community members connect with others is through the disclosure of personal information, as per the value of maintaining interpersonal connectivity (Dholakia et al., 2004). In Stocktwits, investors disclose two types of personal information: demographic information (i.e., real name, bio, URL to more detailed personal information, and location) and trading preferences (i.e., asset type, holding strategy, professional qualifications and risk level). Scholars have concluded that people

disclose personal information to identify with the community and to earn trust (Chaiken & Maheswaran, 1994). Interestingly, Forman et al. (2008) explains how this self-disclosed information is consistent with persistent labeling (e.g., using 'real name') and self-presentation (e.g., revealing location or personal profile). In addition, self-disclosure may even aid hyperpersonal communication where in comparison to Face-to-Face communications; an optimized and manipulated self-presentation is possible (Walther, 1996).

Engage with Others (Social Enhancement).

The need to be accepted and identified as the group is as salient online as it is offline (Ridings & Gefen, 2004). This is particularly visible in stock microblogging where the community is focused on exchanging investing information with high frequencies of interactions among its participants paralleling daily market dynamics, as per the group referent value of social enhancement (Dholakia et al., 2004). Specifically, individuals participate in sending retweets, mentions, and reply tweets to engage others in stock conversations. I examine the different tweet types (retweets, mention and replies) in a typical scenario in the next few sections.

Retweets

Assume investor A has information that the price of Google will be on the uptrend and bought a position. A has eight followers in his/her network and sends a public tweet to them ("$Google on the uptrend bought some"). Any public tweet (denoted by Pu) sent by A is broadcasted to all A's followers (A_1 to A_8). As per Figure 2, A_3, upon receiving

this message from A, and noticing the same Google uptrend retweets A's public tweet to A_3's own followers (B_1 to B_3) with the following content "RT @A $Google on the uptrend bought some." Retweeting (abbreviated 'RT'), an act of forwarding a tweet posted by another, is motivated by the exchange of information, developing interaction with others, and maintaining emotional ties already formed (Zhu & Chau, 2012). It also indicates the ability of the original sender to generate content with pass-along value (Cha et al., 2010), which implies novelty, quality, frequency and the resonance and influence of the message with those of others in the community (Romero et al., 2010). Retweeting is popular as a reliable measure for diffusion and influence in microblogging (Cha et al., 2010; Kwak et al., 2010; Zhu & Chau, 2012). One caveat to note, however, is that since RT is a pass-along action, users who pass along the tweet are credited, while the original source of the tweet may be ignored.

Boyd et al. (2010) provided the following comprehensive summary about retweet. While retweeting may be seen as an act of copying and rebroadcasting, it contributes to a conversational ecology of shared conversational context. This context is different from the traditional conversation structure where groups are bounded in space and time. The traditional conversations derive order from turn-taking and reference to preceding statements, but when the conversation is distributed across a noncohesive network (such as in microblogging) in which the recipients of each message change depending on the sender, these conversational structures are missing. The result is that, rather than participating in an ordered exchange of interactions, people instead loosely inhabit a multiplicity of conversational contexts at once.

Replies

Continuing the scenario from above and assuming that the price of Google did surge for a short period of time, A3 gratefully send a reply 'thank you' tweet to A ("@A Thanks for the tip u made me rich $GOOG"). The sending of reply tweets is another sign of attributing recognition and influence as it usually consists of asking for investing information thus valuing the opinions of the intended recipients. An example of this is when one of A3's followers, B2 is curious and asked a follow up question to A ("@A how long is this pop going to last? $GOOG") directly via a reply tweet. This is illustrated in Figure 3.

Mentions

Mentioning is the action of including others in the on-going conversation (Cha et al., 2010) by citing usernames of the included individuals. Each mention tweet is sent to all the followers as well as those mentioned in the tweet. In this example, A3 publicly praised A again by mentioning A in A3's tweet ("shoutout to @A great observation on $GOOG") Figure 3. Mentioning another can be in public, reply and RT tweets. Mention is a perceived sign of recognition or influence of the mentioned authors. Another type of content message is the inclusion of hyperlinks (usually in the form of shortened urls) linking to more analysis or richer investing information elsewhere. Following this example, A later tweeted a public tweet to his/her followers ("see my analysis of $GOOG http://t.co/5FNWGKmn"). I further illustrate these different types of tweets with respective implications in Table 1.

Identifying Measures of Peer Influence in Stock Microblogging

The discussion about engaging activities in stock microblogging naturally leads to the discussion of identifying and measuring peer influence. Scholars have stated that peer influence is difficult to identify due to endogeneity and correlated effects (Aral & Walker, 2012; Iyengar et al., 2010; Manski, 1993). Iyengar et al. (2010) assert that peer influence should be observed at the individual level and not assumed. It should also cause a change in behavior of peers (Aral, 2011). In addition, peer influence should be based on activity (e.g., visit, retweets) and not pointers (e.g., follower, membership in address book) (Goldenberg et al., 2009). I thus adhere to these guidelines in measuring peer effects using observable social interactions (i.e., retweets, mentions and reply) among online investors in stock microblogging similar to action logs as mentioned in Goyal et al. (2010). However, retweets, mentions and reply tweets are bi-directional, defined as indegree and outdegree actions. Sending these tweets (outdegree) builds relationships in attributing recognition to peers, while receiving them (indegree) is a sign of receiving recognition of peer influence. These are two different types of peer influence measures. As shown in the Figure 4, while outdegree implies effort from the sender in engaging other investors (from A), indegree implies recognition attributed to the receiver by others in the community (to B). Table 2 describes each type of indegree and outdegree measure.

Sentiment Analysis

Stock microblog's message limit of 140 characters is advantageous as it forces users to write succinctly using meaningful keywords. This leads to low cost of information processing and high frequency of generated postings (Java et al., 2009). It

further increases the density of useful information and those keywords are more likely to be repeated by others. One key piece of such information is the investor sentiment or opinion extracted from each microblog. Since these sentiments are derived from text and are not provided by the authors, they have to be either manually or automatically extracted. Two examples of microblog postings with bold sentiment words are shown below:

Bullish posting: $CLW nice **breakout** this am.

Bearish posting: **Shorting** $Amzn 300 pieces @ 131

The first posting indicates a buy (bullish) sentiment with keywords “breakout” while the second indicates a sell (bearish) sentiment with the keyword “shorting”. I used keywords such as these to identify the sentiment of each posting. Lists of common bullish and bearish keywords are shown in Table 3.

Due to the overwhelmingly large volume of microblog posts, it is impractical to label each post manually. Thus I resort to systematic labeling (Das & Chen, 2007). In doing so, I first manually label a set of postings to three distinct sentiments, 1 for bullish, -1 for bearish, and 0 for neutral sentiment--referred to as manual labels. These manually labeled postings are used as the gold standard to evaluate the result of system labeling process on the remainder of the postings. Subsequently two sentiment labeling approaches are attempted 1) lexical knowledge (Kim & Hovy, 2006) and 2) bag of words (Schumaker & Chen, 2009) approaches to automatically label the set of remaining postings which is referred to as system labels.

Lexical Knowledge Approach

One inherent challenge with sentiment classification is domain consideration. Since an identical phrase may have different sentiment in different domains, classifiers trained on one domain and tested on a different domain may not perform well (Pang & Lee, 2008). In this study, I use a domain specific approach with two lists of keywords that pertain to stock investing domain that are commonly used to describe bullish and bearish sentiments. This approach is similar with Das and Chen (2007), known as lexical knowledge approach, in establishing a manually crafted lexicon of hand-picked collection of stock investing words or seeds where each word is pretagged as bullish and bearish sentiment (see Table 3). I then proceed to apply five algorithms: 1) general scoring, 2) weighting, 3) ticker, 4) questions and 5) WordNet to score each posting for bullish and bearish words that are itemized on the two keyword lists. Then an overall score is obtained to determine the final sentiment of each posting.

General Scoring

I begin with two keyword lists: a bullish list and a bearish list of stock investing seed words (see Table 3). These lists are assembled from the consensus of three graduate level students well versed with stock investing jargon. The system compares each word in each posting to both list, and if a match is found, the word is given the following score, 1 for bullish and -1 for bearish. As each word in the posting is processed, a cumulative overall score is calculated. Once all the words in the whole posting are processed, the total score determines its overall bullishness or bearishness. A score > 0 is bullish, < 0 is bearish, and 0 is neutral.

Weighting Approach

This concept is from Hu and Liu (2004) who identify opinion words (adjectives) with respect to features (nouns or noun phrases). An example from their study regarding product review of cameras is the phrase “clear pictures” where “clear” is an opinion word of the feature “pictures”. In this study, I assume the ticker symbol to be a feature and words in the vicinity of the ticker symbol as opinion words. The system selects three words before and after the “<>TICKER<>” symbol and assigns a bullish or bearish score for each word based on the words in the keyword lists. The probability of these words revealing a meaningful and relevant sentiment is much higher than words elsewhere in the posting. Since these scores are in addition to the general scoring, these words add additional weight. The following examples provide clear illustration. Note that the words “pumping” and “long” add to the bullish scores for their respective postings.

Example 1: “thanks to you guys who **have been pumping** <>TICKER<> **the last two** days”.

Example 2: “**Adding** <>TICKER<> **to long list** if it breaks <>DOLLAR<>.”

Postings with Ticker Symbol are Bullish

Postings with just a ticker symbol without any other words are assigned a bullish sentiment. Although such postings do not have any opinion words, I postulate that they are not neutral but have a bullish bias. As explained by Zhang and Swanson (2010) “Self disclosed hold sentiment conveys an optimistic opinion and significantly differs from neutral”. They further assert that when online investors are involved in the discussion on

a particular stock, most of the time they are holding positions in that stock (Zhang & Swanson, 2010).

Questions Are Neutral

Postings with question marks "<>QUESTION<>" are labeled neutral since a question usually does not indicate any sentiment or opinion. I assert that a person who is seeking information is not providing an opinion or making a prediction.

Using WordNet

WordNet (http://wordnet.princeton.edu) is a large lexical database of English nouns, verbs, adjectives and adverbs. Following Hu and Liu (2004), I used WordNet as a tool to expand the existing features by increasing the number of words that are synonymous to the words in the postings. Specifically, for each word *(w)* in the posting, the system obtains all possible synonyms from WordNet. If any of these synonyms are on either the provided bullish or bearish keyword list, then *w* is scored accordingly as being bullish or bearish.

I test the following models:

1. All algorithms (without WordNet)
2. Only ticker words
3. Only all words
4. Both 2 and 3
5. Add Question to 4
6. Add Ticker only to 5

7. Add WordNet to 6.

See Figure 5 for results.

Scenarios 1 and 6 have the highest F-measure score of .596. This indicates that all the features have some influence on the determination of the sentiment of the postings.

Bag of Words Approach

Bag of words is a de facto text analysis technique for financial articles due to its simple nature and its ability to produce suitable vector representation of the text (Schumaker & Chen, 2009). I create a feature vector of all the words in the manual labeled postings set. I then classify this dataset using Naïve Bayes Multinomial classifier. Refer to results in Table 4. The F-measure obtained is .85, with precision .844 and recall .855, a significant improvement as compared to the previous Lexical knowledge approach. Refer to confusion matrix in Table 4.

Due to the challenging nature of sentiment classification, results from the two approaches are highly encouraging as accuracy range of between .6 to .7 is considered good (Das & Chen, 2007; Pang et al., 2002). Since the Bag of Words approach has the best results, I opted to use this approach to automatically assign labels for the remainder unlabeled postings.

Table 1. Type and meaning of a tweet

Tweet type	**Special character(s)**	**Example**	**Meaning**
Retweet	RT	**RT** @etfdigest $GS Goldman Sachs execs in high places of power	Indicates recognition and peer influence of the original author(s) in generating content with pass-along value (Cha et al., 2010) by being included in on-going tweet conversations.
mention	@username in the content of the msg	**@biggercapital @spyder_crusher** $AMZN may be many sold but still no profit	Indicates the ability of the mentioned author(s) to obtain recognition and exert peer influence (Cha et al., 2010).
Reply	@username in the beginning of tweet	**@divtastic** how are u trading $Dell into Earnings? I'm loading on call side put side eow.	Indicates recognition and peer influence that author(s) attribute to recipient(s).
Public	NONE	$GOOG is a disaster	Default tweet to author's group of followers
URL	http	$CVX **http://stks.co/tVV** failed brkout blew thru gap support	Indicates effort of the author in sharing information beyond the tweet (Romero et al., 2010).

Table 2. Incoming and outgoing peer influence measures

Variable Type	**Description**
RT_in	Indegree counts of retweets by other individuals of this individual's tweets
RT_out	Outdegree counts of retweets by this individual of other individuals' tweets
Reply_in	Indegree counts of replies from other individuals to this individual
Reply_out	Outdegree counts of replies from this individual to other individuals
Mention_in	Indegree counts of mentions by other individuals about this individual
Mention_out	Outdegree counts of mentions by this individual about other individuals

Table 3. List of keywords for sentiment analysis

Bullish words			
Up	Upward	uptrend	high
Highs	Long	buy	bought
Buying	Bull	bullish	love
Like	Awesome	break	breaking
Out	Purchase	good	nice
Solid	Blew	expected	happy
Wonderful	Strong	stronger	high
Higher	Strength	positive	covered
Put	Puts	above	in
Off	Bounce	cheap	buyout
Buck	Bucking	cover	covers
Claim	Reclaim	fly	great
Bearish words			
Bad	Worse	worst	sucks
Suck	Lame	dumb	short
Sell	Sold	bear	bearish
Hate	Loss	tank	sad
Sink	Sinks	downward	downtrend
Down	Lower	low	weak
Bye	Negative	under	out
Less	Drop	dropping	kill
Killed			

Table 4. Bag-of-words Naïve Bayes
multinomial classification results for manual postings

Confusion Matrix			
	-1	0	1
-1	1870	74	154
0	35	1470	401
1	8	392	2702
Feature Reduction	Precision	Recall	F
None	.844	.855	.85

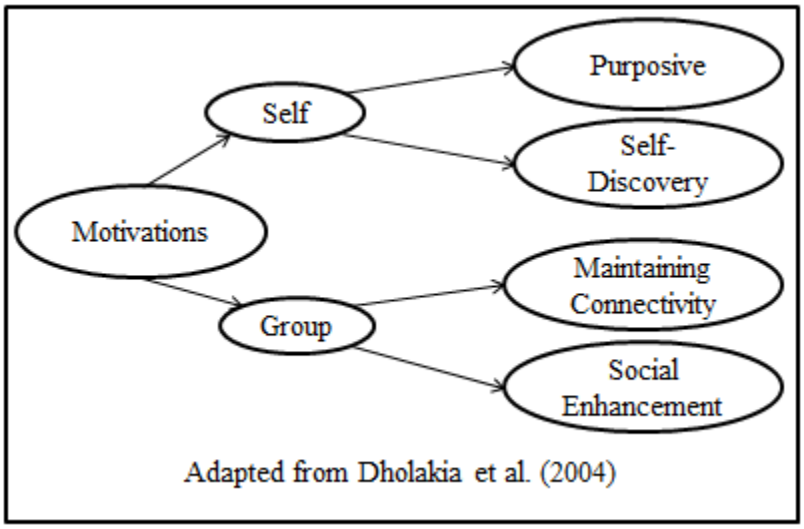


Figure 1. Virtual communities motivation values

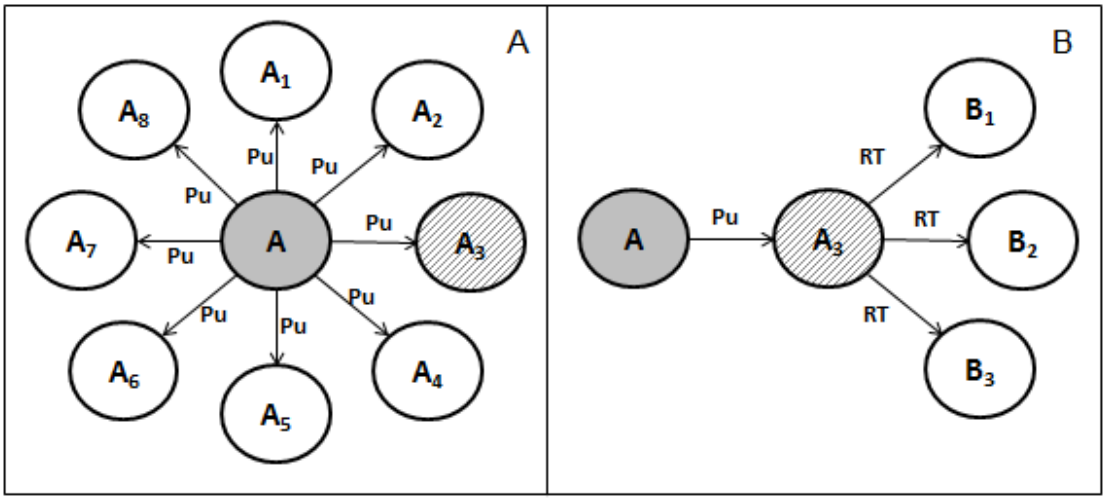


Figure 2. Tweet visibility and reach

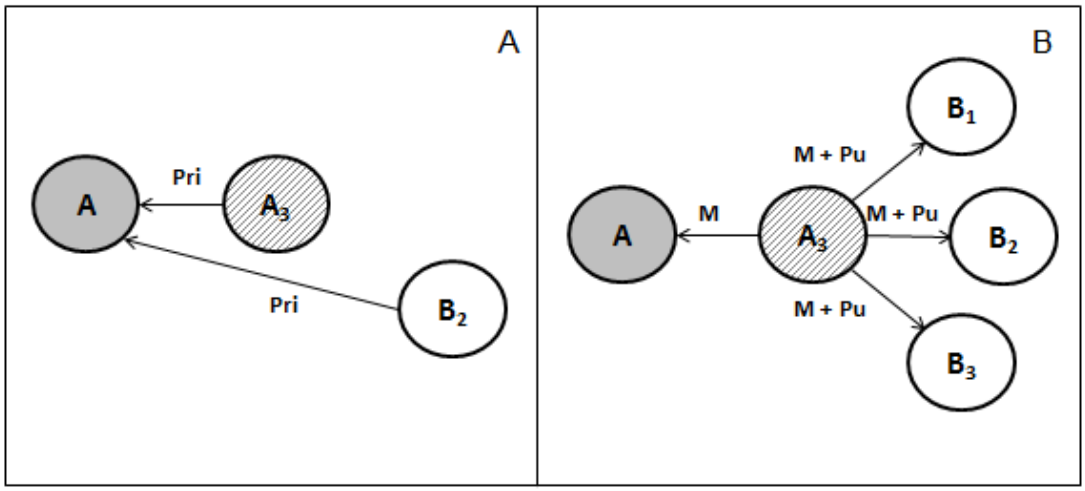


Figure 3. Tweet visibility and reach

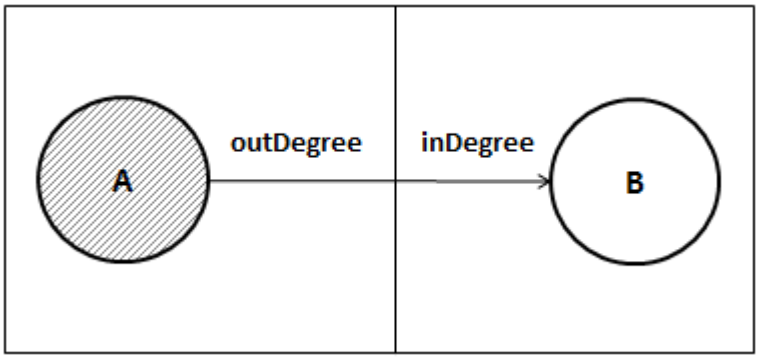


Figure 4. Indegree/outdegree measures

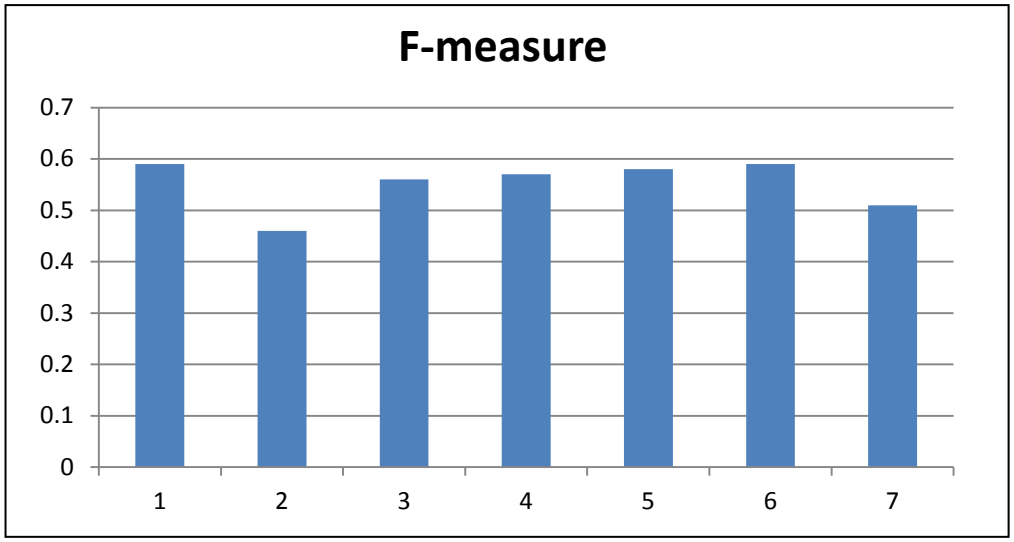


Figure 5. Lexical knowledge F-measure results for manual labeled postings

CHAPTER 3

INDIVIDUAL CHARACTERISTICS, PEER INFLUENCE AND INFORMATION QUALITY

Introduction

Virtual Investing Communities

Prior to the internet, individual stock market investing was a lonely vocation as investors generally worked alone and rarely interacted with each other. About a decade ago, Barber and Odean (2001, p. 41) noted that the internet "is changing how information is delivered to investors and ways in which investors can act upon those information." Essentially, it "alters the way that investors invest, trade, acquire and share information" (Zhang & Swanson, 2010, p.97) by bringing people together in virtual investing communities. More recently, with the advent of social media and Web 2.0 (Ulrich et al, 2008) VICs such as Yahoo Finance and Motley Fools are publishing relevant and valuable financial user generated content (UGC) such as investment recommendations and proprietary analysis, as well as providing an environment where investors can collaborate and discuss, monitor what others are doing, and seek fellowship (Wasko & Faraj, 2005). Some even claimed that sophisticated research and information shared in these communities have the power to move stock prices (Time, 2009). In fact, 55% of investors in public stock-related chat room make profits after transaction costs (Mizrach

& Weerts, 2009). Naturally VICs are popular research channels with active participation from scholars in finance (Antweiler & Frank, 2004), economics (Zhang & Swanson, 2010), and information systems (Chen et al., 2009; Das & Chen, 2007) among others.

Peer Influence

Financial market uncertainties and risk of adoptions drive investors to seek the opinion of others in their decision making process (Becker, 1970; Cancian, 1979), introducing the notion of peer influence (Aral et al., 2009). Peer influence, also known as social influence (Putnam, 1993; Rice et al., 1990), social contagion (Iyengar et al., 2011; Susarla et al., 2012) or peer effects (Bandiera & Rasul, 2006; Sacerdote, 2001), refers to a phenomenon whereby an actor's decision on the adoption of a new product (or behavior) is dependent on other actors' attitudes, knowledge, or adoption (Susarla et al., 2012; Van den Bulte & Lilien, 2001). Scholars have examined peer influence in various context such as marketing (Richardson & Domingos, 2002; Trusov et al., 2009), politics (Dreznar & Farrell, 2004), and blogs (Agarwal et al., 2008). Surprisingly, despite much volume of peer influence inquiry, research efforts on peer influence in the financial domain remain notably absent. This is most likely related to the difficulty in obtaining empirical data on social interactions. The role of peer influence as a parsimonious signal, especially in relation to investor behavior, is important due to its direct impact on the adoption behaviors of investors (Zhang, 2009) amid information overload, a common phenomenon in online channels (Brynjolfsson & Smith, 2000).

Information Quality

Information Quality has recently become a critical aspect in organizations and, consequently, in Information Systems research (Batista & Salgado, 2007). Primarily, information quality describes the quality of the content of information systems (Wang & Strong, 1993) whereas specifically, information quality is defined as accuracy, meaningfulness and timeliness (DeLone & McLean, 1992) and accuracy, comprehensive, currency, reliability and validity (Taylor, 1986). Although information quality research has primarily been a focus of the Information Systems discipline (Wang & Strong, 1993), much work in information quality has also been conducted in accounting and finance (Biddle & Hilary, 2006; Veronesi, 2000).

The Relationship between Peer Influence and Information Quality

Research on peer influence naturally begs the question "do influential individuals have high information quality?" The answer to this question, however, may not be as clear-cut as one presumes. There are two conflicting paradigms in play. On one hand, the purely economic perspective states that rational decision makers seek to interact non-discriminately and for the sake of information alone (Birchler & Butler, 2007; Gu et al., 2007). This infers the presence of high quality content from those deemed influential. On the other hand, due to the effect of social identity (Tajfel, 1978), social exchange (Emerson, 1976), source credibility (Chaiken & Maheswaran, 1994), homophily (Chen et al., 2009; McPherson et al., 2001) and investor overconfidence (Barbaris & Thaler, 2003; Barber, 2001), information quality of exchanged content from influential sources may be negatively impacted. Although this intersection between peer influence and information

quality is interesting and relevant to scholars and practitioners alike, it is largely unexplored. This is probably due to lack of empirical data and the assumption that influential sources inherently possess high information quality.

The following anecdote is helpful in illustrating this notion. Let us assume there are three individuals in an arbitrary network: Andy, Cindy and David. David is interested in investing in the stock market and had conversations with Andy and Cindy about his intention. Andy recently bought a position in Google. A while later David bought Google stock. In the current social influence literature it is logical to conclude that Andy influenced David in David's decision to buy Google. Let us suppose that this is NOT the case and David was actually influenced by Cindy who recommended Google to David in their conversation while Andy did not mention Google at all in his comment. This is a gap in the current social influence literature where the information being exchanged among nodes are ignored due to its unobservable nature or for the convenience of the explanatory model. The relevancy of the topic being discussed, hence the quality of the exchanged content, does have an impact on each individual's influence. This missing piece, measured by an information quality metric, will capture the nuance of information exchanged between individuals and thus will provide a more precise representation of the effect of peer influence.

Research Objectives

I propose an approach to examine the relationships between peer influence and information quality in the context of three dimensions of individual characteristics of stock microblogging: 1) self-disclosed personal demographics and trading preferences, 2)

recognition and 3) effort. As a baseline, I first seek to determine which set of individual characteristics is more salient towards peer influence and which is more salient towards information quality. Then I seek to determine the relationship between individual characteristics, peer influence and information quality. I conjecture that stock microblogging with its features of high volume, succinct, rapid and real-time postings may make more salient factors that tie the relationship between peer influence and information quality.

Based on these discussions I put forward the following research questions:

1. Which individual characteristics are salient towards peer influence?
2. Which individual characteristics are salient towards information quality?
3. How do peer influence relate to information quality?

Contributions

This study contributes to our understanding of the phenomenon in four important ways. The first contribution is to extend the peer influence research with information quality to improve the understanding of the peer influence model. Information quality is a critical factor of peer influence research since the primary motivation for online participation is to seek and share information (Wasko & Faraj, 2005). Even though scholars have examined individual characteristics (e.g., Aral, 2011; Iyengar et al., 2010) and network characteristics (e.g., Goldenberg et al., 2009) in relation to peer influence, the scrutiny of information quality of exchanged information is largely ignored, probably due to empirical constraints. Thus the inclusion of information quality may sharpen the understanding of peer influence since intuitively those deemed with higher information

quality should be more influential. Although the context of this study is in relation to stock market performance and economic outcomes, I assert that the presented models are generalizable to other research disciplines involved in the microblogging domain such as those of movie, product and company tweets.

The second contribution involves extensively exploring various peer influence measures related to information quality. Essentially I extend the current popular peer influence microblogging measures of retweets, mentions and replies (Cha et al., 2010) by incorporating the dimensions of reciprocity and normalization which provide a shaper representation of peer influence. Reciprocity accounts for outdegree of the same peer influence type (e.g., residual of indegree and outdegree retweets) while normalization accounts for the effort of sending total tweets (i.e., residual/total tweets). In addition, I use the individual characteristic measures of self-disclosed demographics and trading preferences, recognition and engaging effort to determine which traits are more salient to information quality versus those that are more salient towards peer influence. These assertions are based on the social identity theory (Ellemers et al., 1999), information processing theory (Chaiken & Maheswaran, 1994) and social exchange theory (Emerson, 1976).

The third contribution of this study explores peer influence research in the online stock investing community, particularly in the nascent stock microblogging channel. I show evidence of peer influence in the stock market, an environment that is easier to quantify than previous studies of peer influence such as adopting Facebook application (Aral & Walker, 2012), prescription drugs (Iyengar et al., 2010) and online participation (Dholakia et al., 2004). I examine investor interactions at a granular level, observing

actual information exchanged between investors, discerning peer influence, evaluating information quality and examining bullishness of sentiment. In essence I support the behavioral finance literature by providing evidence that sentiment of irrational investors do impact stock market, thus adhering to the tenets of investor behavior (Daniel et al., 1998) due to psychological biases such as overconfidence (Hirshleifer, 2001).

In addition, on the practitioner front, I present new knowledge to both platform managers and individual investors. Aral (2011) noted that as vast majority of data available to firms and governmental organizations are observational, making the improved understanding of causal peer influence estimation in such data critical. Better monitoring of peer influence and information quality allude to monetization opportunities for VIC managers and outline mechanisms for investors seeking influential or high quality peers. These different types of peer interactions such as retweet, reply, mention and following unveil different aspects of peer influence (e.g., retweet relates to influence of content while mention relates to influence of the source) and provide a richer set of information for better financial decision making.

Related Literature

In this section, I peruse existing literature on peer influence and information quality in identifying the research gap for this study.

Peer Influence

In the utilitarian perspective, peer influence is defined as “how the behaviors of one’s peers change the utility one expects to receive from engaging in a certain behavior

and thus the likelihood that one will engage in that behavior" (Aral, 2011, p. 2). Peer influence originates from mutual interactions among peers and thus represents the process by which a person's attitude, value, opinion, or decision is influenced by frequent communications with others (Erickson, 1988). Undoubtedly a popular research topic, peer influence has captured the interest of scholars from various disciplines such as sociology, marketing, economics, information systems and network studies. I identified three research streams, namely: 1) examining relationships between various aspects of peer influence with future outcomes such as adoptions of products and services (e.g., Aral & Walker, 2011; Iyengar et al., 2010), 2) identifying opinion leaders (e.g., Aral & Walker, 2012; Goldenberg et al., 2009; Watts & Dodds, 2007) and 3) examining different measures of peer influence (e.g., Agarwal et al., 2008; Iyengar et al., 2010).

Peer Influence with Future Outcomes

Various characteristics of peer influence have been examined by scholars as they relate to future outcomes. For example, Iyengar et al. (2010) correlate social influence (exposure to prior adopters, in-degree nominations from peers and self-reported opinion leadership) with adoption of a new prescription drug among physicians. They confirmed the saliency of peer influence even after controlling for marketing effort and high usage volume. In addition they also found that indegree nominations from peers is positively correlated with adoption and that it is a different type of social influence construct from self-reported measures, and high volume users are more influential than light users. Aral and Walker (2011) conducted a randomized experiment in Facebook viral messaging, correlate peer influence (number and percentage of adopters in the network) and

individual characteristics with adoption of Facebook application. They demonstrated how randomized experiments can be used to identify peer influence effects in networks and the significance of product viral features in increasing those effects. A surprising discovery was that passive messaging leads to more social influence although active messaging relates to more engagement and sustained product use. Susarla et al. (2012) examined how different social influence mechanisms such as conformity, social learning and the role of opinion-makers impact the demand and diffusion of YouTube content. They examined social network features of indegree and outdegree centrality of friends' networks (conformity), connections to peers outside the friends' networks (social learning), degree centrality of subscribers' network (opinion-makers) and product characteristics. They found that actors connected to others outside his/her network are more influential (social learning) than those connected to peers in his/her friend networks (conformity). And early adopters are pivotal in persuading others to adopt. Katona et al. (2011) relate three social influence factors: local network structure, characteristics of adopted peers and characteristics of potential adopters, to adoption probability of a European social networking site. They discovered that those who are highly connected to many adopters have a greater adoption probability. And the density of connections in a group of adopted consumers has a strong influence on adoption probability of an actor. In addition network position and certain demographic variables are good predictors of adoption as well.

From prior discussion, I note that the extant literature has confirmed the saliency of peer influence in relating to future outcomes. However I note that there is a paucity of

iterature in relating peer influence to the information quality of exchanged information among individuals in the network.

Identification of Opinion Leaders

The next literature stream examines identification of opinion leaders. This is known as the "influentials" hypothesis (Valente, 1995) , a well-accepted notion that opinion leaders exist and that they are catalysts for promoting diffusion of opinions, innovations and products (Aral & Walker, 2012; Coleman et al., 1957; Rogers, 2003; Valente, 1995; Van de Bulte & Joshi, 2007). Early studies of influentials focused on identifying structural network measures such as degree or betweenness. One such is Goldenberg et al. (2009) who identified two types of hubs (individuals with large degree of links to peers): the innovator and the follower hubs. They concluded that hubs in general tend to adopt earlier in the diffusion process. And although innovative hubs have a greater impact on the speed of the adoption, follower hubs have a greater impact on the total number of adoptions. Identification of hubs aid buzz marketing, leading to faster growth and increased market size.

With the prevalence of UGC there has been an explosion of consumer characteristics related to personal interests, preferences, behaviors and product tastes (Aral & Walker, 2012) and such details should be included in the peer influence model to enrich the understanding of these relationships. For example Aral and Walker (2010) correlated individual demographics, school and employment history, product tastes and social participation with adoption. They found that individual characteristics such as gender and relationship status are significant. On the contrary, some scholars diminish the

importance of influential. One such is Watts and Dodds (2007) which discovered that large cascades of influence are not driven by influential individuals but by a critical mass of easily influenced individuals. This finding downplays the importance of influential and places the focus on the masses, in identifying and harnessing the right audience. In a more recent study, however, Aral and Walker (2012) concluded that influential and susceptible groups play a role in the peer-to-peer diffusion and each is distinct from the other, and an individual is not likely to be in both groups. The influential group is larger, thus targeting should focus on the attributes of current adopters instead of their peers, the adoptee.

Similar with literature in the previous section, I note that current research in influentials and susceptible investigate network structure and individual characteristics but rarely examines information quality of exchanged information among individuals in the network.

Measures of Peer Influence

In this section I review different measures of peer influence examined in past literature. For example, Agarwal et al. (2008) applied various network based parameters such as inlinks, outlinks, comments, and length of blog posts to examine influence and thus identify top community bloggers. They concluded that influential bloggers are recognized by peers, can generate follow-up activities, have novel perspectives or ideas and are often eloquent.

A popular set of measures is the exposure to adopted peers such as number or percentage of adopted peers in the individual's network of ties (Aral & Walker, 2011;

Iyengar et al., 2011). These measures then led to development of social network measures of degree centrality and betweenness centrality, which are very popular in the peer influence literature. For example, Susarla et al. (2012) used degree centrality measures of direct ties of links from peers within the friend group, subscriber group and external group relating to adoptions of YouTube content. Katona et al. (2011) used degree and betweenness centrality to measure network effect of peer influence in a social network site. It is common for such ties or links to be based on pointers (e.g., being in the same friend network) and not based on activity (e.g., recommendation, conversation, citation). Scholars cautioned against the assumption that such links infer influence. According to Aral (2011) highly central individuals or individuals or high degree are not necessarily influential as individuals must cause behavioral change in the network rather than just simply being connected to his or her peers. Goldenberg et al. (2009) also stated that a link needs to be defined by activity, not by pointers as a link between two people in a social networking site does not necessarily imply influence.

Peer influence measures via identification of opinion leaders are obtained from three sources: survey (self-reported or self-perceived), sociometrics (observations of individual behavior) and key informants (Iyengar et al., 2011). Whereas self-perceived is a more popular technique with marketing academics, sociometric is more popular among social network analysts. It is likely that self-reported is biased upwards and that it reflects self-confidence rather than actual influence (Iyengar et al., 2001). Sociometric measures such as indegree and/or outdegree nominations from/to peers (Iyengar et al., 2011) are less popular due to empirical constraints. But these are the more reliable measures since

being nominated by peers as someone they turn to for expertise or discussion is likely to be a true source of influence (Iyengar et al., 2011).

From the review I note that there is no attempt to measure information quality to better understand its role in the peer influence research.

Measures of Peer Influence in Microblogging

In this section I specifically focus on measures of peer influence in the microblogging domain, which have been examined quite extensively despite microblogging's infancy due to the popularity of Twitter. Scholars have ascertained that microblogging is a viable area for viral marketing, customer relationship management, and eWOM branding initiatives (Jansen et al., 2010; Milstein et al., 2008). This nascent yet active research stream infers to the relevancy and richness of user interactions in microblogging communities. Literature in the diffusion of information and influence in microblogging has been interesting and relevant. Let us peruse a few notable examples focusing on the relationships being investigated and measures used.

Based on use and gratification paradigm and social identity theory, Zhu & Chau (2012) examined how different psychological states and the interaction between them impact message forwarding or retweeting for a firm. The authors concluded that information overload (followers), interest in communication topic (mentions), desire to participate (number of tweets posted), and self-identification (replies) significantly impact whether the firm's tweets are retweeted. They found that while information overload has a negative correlation with retweet, the other three states are positive. In summary the authors have shown that measures of microblogging are driven by

psychological states. Similarly, Boyd et al. (2010) concluded that retweeting is a medium to involve disparate participants into a conversation and allow a fast-pace conversational environment to emerge.

Cha et al. (2010) compare the three measures of social influence: number of followers (indegree), retweets and mentions. They found that follower is related to popularity and not influence while influence is a result of concerted effort of engagement and knowledge contribution. In addition, they also concluded that retweets are driven by the content value of the tweet while mentions are driven by the name value of the user. Kwak et al (2010) found rankings of influential by followers and page rank to be similar but rankings by retweets to be different from the first two, indicating the difference in information diffused in the two types of measurements. This is in support of Cha et al. (2010) in uncovering the difference between the two measures. Weng et al. (2010) proposed TwitterRank, an extension of PageRank, based on networks of follower-followings to measure the influence of users in Twitter. They also found that reciprocity can be explained by the presence of homophily and not due to information diffusion. Although past literatures in microblogging apply peer influence measures but few have attempted to provide any theoretical explanation for them. In this study I review and explain the theories behind these measures in relating to information quality.

From the literature review I note that although much work has been done in understanding peer influence with future outcomes, identifying opinion leaders and investigating measures of peer influence, there is an absence of literature investigating the relationship relating individual characteristics and peer influence with information quality. This is an interesting and relevant research gap because information quality is

such a significant factor in social exchanges and is especially poignant in the finance and stock investing areas.

Information Quality

Information quality describes the quality of the content of information systems (Wang & Strong, 1996). Among the many characteristic definitions of information quality, two are well quoted. The first is from DeLone and McLean (1992) - “accuracy, meaningfulness and timeliness” and second is from Taylor (1986) - “accuracy, comprehensive, currency, reliability and validity”. The focus of information quality is driven primarily by each field of research. For example, IS scholars focus on role of information quality in information quality framework and measures (Aladwani & Palvia, 2002; Srinivasan, 1985; Wang & Strong, 1996), technology acceptance (Ahn et al., 2007; Lederer et al., 2009) and web quality (D’Ambra & Rice, 2001). Meanwhile, in finance and accounting, researchers focus on the role of information quality as a critical component of decision-making as used by investors or stakeholders to update projections of future growth rate, inflation rate, and interest rate, and in turn how information quality impacts investor expectations on stock market prices (Veronesi, 2000). For example information quality is generally examined in the context of accounting information (Biddle & Hilary, 2006), financial reporting/information disclosure (Hermalin & Weisbach, 2012; McDaniel et al., 2002) and stock returns/asset pricing (Brevik & d’Addona, 2010; Epstein & Schneider, 2008; Veronesi, 2000). A few scholars have initiated effort in understanding information quality in VICs (Sprenger & Welpe, 2010;

Zhang, 2009) made possible via enormous high volume investor interactions in nascent UGC channels such as stock microblogging and stock message board postings.

The notion of information quality is highly subjective, as it varies considerably among users, depending on how it is utilized. Wang and Strong (1996) proposed an information quality framework that covers four core dimensions of information quality: intrinsic, contextual, representational and accessibility. Each of these core dimensions captures a respective aspect of data quality. In line with the notion that a high degree of accuracy increases its objectivity (DeLone & McLean, 1992), I distinctively select an objective measure of "intrinsic information quality" from Wang and Strong's (1996) framework -- author average accuracy in predicting future stock price outcomes. The authors further defined intrinsic information quality with four factors: accuracy, believability, objectivity and reputation (Wang & Strong, 1996 p 20). This parallels a similar four factors of intrinsic information quality by Delone and McLean (1992): accuracy, precision, reliability and bias-free.

Peer Influence and Information Quality

Although peer influence has been extensively studied in many disciplines, I note that there is a lack of research effort in understanding the relationship between peer influence and information quality. I seek to fill this gap by investigating a nascent VIC channel, stock microblogging, in the financial stock investing domain, a domain that is interesting and relevant to both scholars and practitioners alike. Intuitively one would posit that with stock microblogging's smaller sized, stock investing specific VIC channel, community members are able to encourage social bonds that elucidate trust and

knowledge contribution resulting in the creation of more salient peer influence in the community and thereby higher exchange of information quality.

Theoretical Framework and Hypotheses Development

In this section I discuss the theoretical framework and hypotheses development in this study. As presented in Chapter 2, individuals are motivated to participate in online communities due to four core objectives: *purposive* and *self-discovery*, both self-referent values, *maintaining interpersonal connectivity* and *social enhancements*, and both group-referent values (Dholakia et al., 2004). Similarly Stocktwit investors are motivated to seek and share investing information (*purposive value*), disclose personal information (*maintaining interpersonal connectivity*), gain recognition (*social enhancement value*) and to engage with others (*purposive and social enhancement*). Based on social identity theory (Tajfel, 1978), social exchange theory (Emerson, 1976), source credibility (Chaiken & Maheswaran, 1994), homophily (Chen et al., 2009; McPherson et al., 2001) and investor overconfidence bias (Hirshleifer, 2001) I examine individual characteristics of self-disclosure (*maintaining connectivity*), recognition (*social enhancement*) and engaging others (*purposive AND social enhancement*) in relating to peer influence and information quality. The motivation value of self-discovery is not mapped in this model. I first examine relationships between individual characteristics and peer influence, then individual characteristics and information quality, and finally peer influence to information quality. This conceptual framework is outlined in the Figure 6.

Relating Self-Disclosure with Peer Influence and IQ

Disclosure of self-identity in virtual communities is an interesting research topic. Social identity theory (Ellemers et al., 1999; Tajfel, 1978) asserts that individuals affirm a clear and consistent sense of self and wish to feel connected to others and receive identity-affirming feedback from the community (Forman et al., 2008). In fact individuals are motivated to present their identities in everyday social life (Ma & Agarwal, 2007). Thus they disclose identity and preferences to identify with the community and to be in good standing (Postmes, 2000; 2005; Sassenberg, 2002). By reaching a consensus regarding identities, people feel understood and obtain a sense of continuity and coherence (Ma & Agarwal, 2007; Swann et al., 2000). Such behaviors facilitate formation of relationships, common bonds and social attractions that community members value (Ren et al., 2007). Furthermore, due to the notion of hyperpersonal (Walther 1996), self-disclosed personal information makes us feel even closer than compared to a face-to-face environment as the a hyperpersonal message sender has a greater ability to strategically develop and edit self-presentation, enabling a selective and optimized presentation of one's self to others.

In Stocktwits, investors disclose two types of self-identifiable information. One type is demographic information such as full name, location, bio and URL. The other type is trading preferences such as assets traded, approach, holding period and trading experience. Both sets of self-disclosed information help the community to identify with the particular individual via similar background or similar trading preferences (See Appendix page 169). Investors tend to seek peers with similar portfolio characteristics or with similar trading strategies in decision making. Therefore, individuals who self-

disclose should be rewarded with positive feedbacks from peers resulting in peder recognition and influential effects. With this in mind, I postulate that:

H1A All else equal, demographic and trading preference self-disclosures should positively correlate with peer influence.

Similarly, self-disclosed information may also favorably relate to information quality of trading information. Forman et al. (2008) found that reviewer's self-disclosed identifiable information in Amazon marketplace is used by consumers to supplement or even replace product information in making purchase decisions and evaluating the helpfulness of product reviews. Based on source credibility theory (Chaiken & Maheswaran, 1994; Hass, 1981) the authors stated that attributes of an information source have powerful effects on how people respond to messages. Forman et al. (2008) found that reviews with identity-descriptive information are read more positively and associated with a subsequent increase in product sales. These reviews gain trust and are recognized by community members. Information acquisition is efficient when the expert is identifiable resulting in people perceiving knowledge to be more useful and paying greater attention to it (Ma & Agarwal, 2000).

In addition, identifiable information may increase the level of personal accountability. This feeling of accountability is induced by a sense of copresence (Golfman, 1959). In short, people are more careful and accurate when they feel the presence of others and when they know that others can identify them (Ma & Agarwal, 2007). This phenomenon is even more salient in the Stocktwit community where peers are small in numbers, but frequency of interactions is high. So tweets from those who

self-disclosed should contain higher information quality. With this in mind, I postulate that:

H1B All else equal, demographic and trading preference self-disclosures should positively correlate with information quality.

Relating Recognition with Peer Influence and IQ

Peer recognition systems are known to motivate quality knowledge contribution and online participation (Resnick et al., 2000). By the same token "firm recognized" recognition is also known to have a similar effect (Jeppesen & Frederiksen, 2006). In Stocktwits, expert investors are recognized by the "suggested" label, which is a 'black box' firm recognized status, assigned by Stocktwits. According to Jeppesen and Frederiksen (2006) firm recognitions are also peer recognitions as those recognized by the authority are inadvertently recognized by peers to signal high expertise or competence.

Due to the ease of use of stock microblogging, investors flood Stocktwits with high volume, succinct, rapid and real-time tweets (Sprenger & Welpe, 2010). Information processing literature may shed some light into individuals' behavior in consuming this information. There are essentially two information processing types: systematic and heuristic. Chaiken and Maheswaran (1994) explained that systematic processing (central route) implies that people have formed or updated their attitudes by actively attending to and cognitively elaborating persuasive argumentation (e.g., reading stock tweets). In contrast, heuristic processing (peripheral route) (Chaiken, 1980) implies that people have formed or changed their attitudes by invoking heuristics such as "experts can be trusted,"

"majority opinion is correct" and "long messages are valid messages." Since systematic processing takes effort and is cognitively more demanding than heuristic processing, it is fair to assume that heuristic processing is predominant when effort is overbearing, cognitive capacity is limited (e.g., as in the Stocktwit community) or when time does not permit extensive information processing (e.g., volatile market). Due to the high volume rapid tweets and fluctuating market dynamics, investors in Stocktwits are likely to resort to heuristic cues such as "suggested" label as mental short-cuts, rule of thumb or guidelines (Metzger et al., 2010) to determine credible sources to trust, which in reality may not lead to quality investing information. With these discussions in mind, I postulate that:

H2A All else equal, being on the suggested list should positively correlate with peer influence.

H2B All else equal, being on the suggested list should negatively correlate with information quality.

Relating Engaging Effort with Peer Influence and IQ

Social exchange theory (Emerson, 1976) states that individuals participate in social interaction due to the expectation that it will lead to social rewards such as approval, status and respect (Wasko & Faraj, 2005). It supports the motivations for participating in VIC through publishing stock investing information and engaging with fellow investors. Influential individuals are likely to develop trust, have a strong identity with the community, and to have an obligation to participate and abide by community norms (Nahapiet & Ghoshal, 1998; Wasko & Faraj, 2005). Moreover, these dimensions

spur the overall desire to engage in social exchange (Putnam, 1993). According to Wasko and Faraj (2005) the main objective of participating in virtual communities is to seek and share information known as the purposive value mentioned earlier (Dholakia et al., 2004). It is sought after by individuals who seek informational value in making better investing decisions and for those who seek social enhancements and maintaining interpersonal connectivity. The need to be accepted, to bond, and to be identified with the group is as salient online as it is offline (Ridings & Gefen, 2004). Furthermore, it is a universal norm requiring that aid received from others be compensated (Gouldner, 1960), thus the notion of reciprocity. Reciprocity is a key element in the density of social relationships (Uehara, 1990) whereby individuals feel a need to reciprocate or expect reciprocity in return for sharing information and engaging with peers.

Conversations, in the context of stock tweets, involves discussing alternatives, making predictions, asking questions, reporting observations, contributing opinions, sharing analysis and announcing decisions. Conversation is critical in the contagion of popular ideas about financial markets. People tend to pay more attention to ideas or facts that are reinforced by conversations, rituals and symbols (Hirshleifer, 2001). Furthermore, high volume tweets reduce perceptions of riskiness (Heath & Tversky, 1991) as repeated exposure to an object (or another person) increase familiarity and subsequent liking of the object (Pollock & Rindova, 2003). Perception of riskiness is a significant factor particularly in the domain of financial stock investing. These tend to lead to a positive relationship with peer influence which is not gained spontaneously or accidently but through concerted effort and consistent personal involvement.

In the context of Stocktwits there are two types of effort; first is in the contribution of opinions and investing information, primarily in the form of publishing stock tweets. The second is the effort of engaging with other investors via retweeting, sending reply tweets, and mentioning others in their tweets. I examine the dimensions of out-degree retweets, out-degree mentions, out-degree reply tweets and total posted tweets as different measures of individual effort. In short, those consistently publishing investing information and actively engage with fellow investors tend to be recognized by peers resulting in peer recognition and influential effects. With this in mind, I postulate that:

H3A All else equal, measures of effort should positively correlate with peer influence.

Finance scholars have determined that there are social psychological biases that influence the behavior of investors (Barber & Odean, 2001). One such is the illusion of control (Langer 1975), when people overestimate their ability to control events. Another is the illusion of knowledge (Burger 1984), when people with access to information believe they are more knowledgeable than they really are. This is related to biased self-attribution where people tend to attribute good outcomes to their own abilities and bad outcomes to external circumstances (Hirshleifer, 2001). Both illusion of control and illusion of knowledge lead to the phenomenon of investor overconfidence where people tend to overestimate their own knowledge about a stock (Barber & Odean, 2001; Hirshleifer, 2001) and thus overstating their opinions. In fact scholars argued that investors are more likely to be overconfident about private information (e.g., those obtained from others in the community via interactions) than information that is publicly available (Daniel et al., 1998; 2001). These biases would further impair the existing

uncertainty derived from difficulty and subjectivity in determining the value of stocks (Baker & Wurgler, 2007). Overconfidence thus leads to publishing more tweets, expressing more opinions, engaging more with others and contributing more analysis and predictions. According to Shiller (1999), owing to limited cognitive capacity, people tend to pay much more attention to ideas or facts that are reinforced by conversation, ritual and symbols. This results in the source of such information being more visible and thus deemed influential. In short, overconfidence may lead to higher volume of tweets and higher level of engaging activity but may have a negative effect on the quality of shared trading information. With this in mind, I hypothesize that:

H3B All else equal, measures of effort should negatively correlate with information quality.

Peer Influence and IQ

The peer influence measures focus on two different dimensions of social interactions in the StockTwit community: indegree mentions and indegree retweets. Although both are interpersonal activities in microblogging that focus on an individual's influence in leading others to engage in a certain act (Cha et al., 2010), and are well accepted as measures of social influence (Cha et al., 2010; Kwak et al., 2010; Weng et al., 2010; Zhu & Chau, 2012), they represent different perspectives of the influence of a person. These two measures are indegree mentions and indegree retweets.

First, the indegree mentions centers on individuals who are frequently mentioned by peers in their tweets, focusing on the name value of the person being mentioned in a particular tweet, alluding to the person's ability to engage others in a conversation about

or relating to him or her. It is a norm for the community to rally around and mention individuals who possess high quality trading information. Not only it is an act of recognition, it is also an act of showing appreciation towards these individuals. From source credibility theory I learned that experts possess high credibility (Forman et al., 2008) probably due their hold on information quality that mitigates adverse outcomes from decision making. Similarly, in the stock microblogging community peers do recognize and appreciate experts by citing such individuals (i.e., mentioning them) in their conversations (i.e., tweets). Thus, I postulate that:

H4A All else equal, individuals with a higher peer influence measure of indegree mentions should be positively correlated with information quality.

Second, the in-degree retweet focuses on the individuals with the ability to induce retweets from his or her followers. Since individuals in communities tend to interact with similar others, followers are likely to have the same characteristics with the individuals they follow, leading to the phenomenon of homophily (Chen et al., 2009; McPherson et al., 2001). In Stocktwits these similarities are probably determined by each individual's self-disclosed personal demographics and trading preferences. Homophily is driven by confirmation bias (Brehm et al., 2005) when one has a priori beliefs prior to the interaction, being influenced by those of the same opinions while shunning others of dissimilar opinions. As such, homophily may have an adverse effect on the information quality of trading information being exchanged among such peers.

Furthermore, in the stock community, homophily is further exacerbated due to investor's bias towards diversification of portfolio or familiarity hypothesis (Huberman, 2001). In essence, investors are likely to focus on a few familiar stocks instead of being

diversified in his/her investments. This leads to investing ideas and decisions that are similar and frequently retweeted among followers but may not be of high information quality due to lack of diversity. With this in mind, I postulate that:

H4B All else equal, individuals with a higher peer influence measure of indegree retweets should be negatively correlated with information quality.

Investor Sentiment, Peer Influence and IQ

Another feature from the microblogging author is the sentiment pertaining to each stock ticker in the tweet discussion. Investor sentiment is part of effort in engaging peers in discussing about a particular ticker. It also stems from the motivation of purposive value, which is to seek and share investing information opinions and information.

In models of investor sentiment, uninformed traders based their decisions on various sources of information which in aggregate, although noisy, do influence stock prices (De Long et al., 1990). These opinions and beliefs are captured in the tweet in the form of sentiment. In this section I discuss how investor sentiment relates to peer influence and information quality.

People pay more attention to negative than positive news (Luo, 2007). This is because according to prospect theory (Kahneman & Tversky, 1973), losses weigh more than gains. Furthermore, Chevalier & Mayzlin (2006) echo that an incremental negative review is more powerful in decreasing sales than an incremental positive review is in increasing sales. This is because negative words have more impact and are more thoroughly processed than positive information (Baumeister et al., 2001; Rozin & Royzman, 2001). This alludes that stock tweets with bearish sentiments are more likely to

capture attention, thus correlate with higher peer influence. With this in mind, I hypothesize that:

H5A All else equal, investor sentiment should be negatively correlated with peer influence.

Bearish sentiment is pessimistic or negative news, and it implies a declining stock prices. Bullish sentiment, which is related to overconfidence and over-optimism, on the other hand, often lead to wishful thinking, a phenomenon that is related to speculative bubbles and information mirages (Seybert & Bloomfield, 2009). “Wishful thinking is the formation of beliefs and decision making according to what might be pleasing to imagine instead of appealing to evidence, rationality and reality” (Wikipedia, 2010b). Wishful thinking is highly influential in markets where many traders who each hold a small bit of information have to rely on inferences from observed behavior in order to estimate asset values (Seybert & Bloomfield, 2009). Since VICs possess these characteristics, they are susceptible to wishful thinking. I postulate that since wishful thinking has a negative implication on accuracy, bullish sentiments should correlate with lower information quality than that of bearish sentiments.

Furthermore, market frictions (Miller, 1977) and behavioral biases (DeBondt & Thaler, 1985; Hirshleifer, 2001) may cause price to deviate from fundamentals in the short run and short sellers are exploiting these situations to their benefit through short selling. Short selling is the practice of selling stocks at a higher price with the intention of purchasing them later at a lower price, a practice that involves a deeper level of trading knowledge and acumen. This suggests that short sellers, as a group, are more sophisticated than the average investor (Diether et al., 2008) as they possess higher

trading information. In this study, tweets from short sellers are identified with bearish (negative) sentiments. With this in mind, I hypothesize that:

H5B All else equal, investor sentiment should be negatively correlated with information quality. Table 5 lists the hypotheses summary.

Data and Variables

Data for this study is provided by Stocktwits (http://www.stocktwits.com), a popular stock microblogging channel established since October 2008. Stock interday and Dow Jones index used in generating information quality measures are obtained from Google finance (http://www.google.com/finance). After initial preprocessing removing nonrelevant posts I obtained over 360,000 postings from 8935 authors pertaining to 4570 stock tickers. I first aggregate these tweets at the author-ticker-day level to obtain predictive outcome and then I aggregate to the author-week level.

Measures

Dependent Variables

The two groups of dependent variables in this study are peer influence and information quality. I discuss each group in this hereafter.

Generating Measures of Peer Influence

In measuring peer influence, I first adhere to Goldenberg et al. (2009) and Trusov et al. (2008) by defining links by activity (e.g., retweets, mentions) and not by pointers (e.g., follower) as a pointer between two individual in a social networking site does not

necessary imply influence (Goldenberg et al, 2009). Second, I adopt Iyengar et al. (2010) in using indegree nominations of peers as a measure of peer influence. According to Iyengar et al. (2010), people who are often nominated by peers as someone they turn to for expertise or discussion are likely to be true sources of influence. Based on these two assumptions, I thus propose five groups of peer influence measures: 1) baseline indegree counts, 2) residual, 3) normalized, 4) normalized influence by peer outdegree and 5) normalized influence by unique peers. In the current microblogging literature, scholars have yet to reach a consensus regarding correct measures for peer influence. However, the more popular measures of peer influence are number of followers (Cha et al., 2010; Li & Shiu, 2012; Weng et al., 2010), retweets (Cha et al., 2010; Kwak et al., 2010; Li & Shiu, 2012; Sprenger & Welpe, 2010; Zhu & Chau, 2012), mentions (Cha et al., 2010), and URL propagations (Galuba et al., 2010; Romero et al., 2010).

For the first group, I adopt two basic indegree measures from the literature, namely number of indegree retweets and mentions and tested another measure: number of indegree reply messages. As discussed in Chapter 2, replies are private tweets between individuals.

For the second group I control for the effect of reciprocity as explained by Weng et al. (2010) by subtracting outdegree from the indegree measures resulting in residual measures (e.g. RT_diff = RT_in – RT_out). Agarwal et al. (2008) defined this as InfluenceFlow which intuitively suggests that the more inlinks a blog post acquires the more recognized it is while an excessive number of outlinks jeopardizes the novelty of a blog post. Thus an individual with many indegree accompanied by many more outdegree should have a lower influence compared to another with high indegree but lower

outdegree. Because peers are obligated to reciprocate, sheer number of outdegree naturally begets high number of indegree. Hence using indegree measure without considering outdegree may be misleading. Despite its novelty and intuitiveness, it is surprising that few research studies in microblogging adopt this measure.

The third group controls for the total tweet sent by the individual in relation to his/her residual peer influence score. I normalized each residual score with the number of total tweets posted by the individuals during the same period (e.g., RT_norm = RT_diff/total tweets). The intuition is that two individuals with the same residual score (e.g., 10) might have different levels of peer influence when effort is accounted for (e.g., 10/2 total tweets or 10/10 total tweets). Intuitively the individual with the higher normalized score should have a higher level of peer influence.

The fourth and fifth groups are new measures proposed in this study to measure how relevant an individual is to his/her peers by accounting for the proportion of attention given by the individual's peers, both in terms of peers' outdegree count (group 4) as well as peers' number of peers (group 5). The intuition is that an individual with more peer attention should be more influential. I based this on the TF-IDF (term frequency – inverse document frequency) concept (Salton & McGill, 1989) from information retrieval. Term frequency (TF) (Wu et al., 2008) refers to how relevant a word (term) is in a collection or corpus. Group 4 and 5 are operationalized below.

Normalized influence by peer outdegree (NIPO) = (indegree to A) / (outdegree from A's peers)

Normalized influence by unique peers (NIUP) = (count of A's indegree peers) / (count of A's peers' peers)

An example of *Normalized influence by peer outdegree* (NIPO) and *Normalized influence by unique peers* (NIUP) are provided (See Appendix page 170). Table 6 outlines all the dependent variable measures for this study.

Generating Measures of Information Quality

I adhere to the definition of DeLone and McLean (2004) and Wang and Strong (1996) in selecting an objective measure of intrinsic information quality: a measure comparing the bullishness index of tweets per ticker posted by each individual each day against each ticker's same-day and next day simple return. I find this information quality measure of comparing prediction with stock price movement to be timely, accurate, and complete (DeLone & McLean, 2004). In the VIC literature, this measure was applied in Sprenger & Welpe (2010) and Zhang (2009) to measure information quality in their respective studies. The measure of information quality is defined in Equation 1.

$$\text{Quality} = \begin{cases} = 1 \text{ if } (s_{it}/R_{it}) > 0 \\ = 0, \text{ otherwise} \end{cases}$$

(Sprenger & Welpe, 2010)

(Equation 1)

where s_{it} is the bullishness index from each individual on day t (and day t +1) associated with stock i. And R_{it} is the return of stock i on day t. Those with bullishness index = 0 is removed. Thus *information quality* is 1 when sentiment corresponds with return (e.g. bullish sentiment with positive return, or bearish sentiment with negative return) and 0 otherwise. To measure *information quality* for all stocks per individual, I generate the average.

Independent Variables

The independent variables for this study consist of individual characteristics of investors organized by self-disclosed attributes, engaging effort and recognition measures. I adhere to Iyengar et al. (2010) and Aral and Walker (2010) in using individual characteristics to explain peer influence. See Table 7.

Self-disclosed

Self-disclosed identifiable information includes user demographics (real name, bio, url, education) and trading preferences or experiences (experience, approach, risk, holding and trading styles). This information is disclosed in the profile page of each author. For the sake of parsimony, I assign demographics or trading =1 if any of the enclosing traits exists.

Recognition

Recognition attribute consists of the author being on the suggested list assigned by Stocktwits.

Engaging Effort

Engaging effort consists of lagged outdegree counts of retweets, mentions and replies and total tweets sent. The lagged measures are employed to overcome endogeneity issues highlighted by Manski (1992) and well discussed by many prominent scholars (Aral, 2011; Iyengar et al., 2010). Total tweets sent, a measure of participation, is a prominent measure of influential individuals (Iyengar et al., 2010).

Controls

I control for the following variables in this study.

Lagged Follower and Following

I include lagged follower and following counts to control for the effects of autocorrelation of follower and following from past periods.

Lagged Influence Dependent Variables

I include lagged dependent variables such as lagged indegree RT, mentions and replies to control for the effects of autocorrelation that is common with stock market data (Antweiler & Frank, 2004).

Lagged Information Quality Dependent Variables

I include lagged dependent variables for information quality to control for the effects of autocorrelation of past information quality.

Investor Sentiment

As a poignant factor of any investing community, investor sentiment is part of the effort of individuals in sharing his/her opinions or stock investing advice. In addition I also include disagreement index, a measure of how agreeable is the author's overall sentiment. This is discussed later in this section.

Sentiment Similarity

As stated by Gu et al. (2007), individuals in virtual communities tend to demonstrate homophilous behavior, reinforced by cognitive dissonance and the availability of like-minded peers. In the Stocktwit community there is a strong tendency to follow the sentient of peers. I thus control for the effect of homophily (McPherson et al., 2001) in this study by the following three sentiment similarity measures: sentiment similarity index, sentiment similarity binary index and average sentiment distance, based on sentiment by network type where type is retweet, mention or reply. The details are discussed later in this section.

Generating Measures of Sentiment Bullishness and Disagreement

To generate aggregated metrics for sentiment bullishness for each individual I refer to Antweiler & Frank (2004). See Equation 2.

$$\text{Bullishness Index} = \ln\left[\frac{1 + M^{BULL}}{1 + M^{BEAR}}\right] \quad \text{(Equation 2)}$$

M^{BULL} is the total number of bullish tweets while M^{BEAR} is the total bearish tweets. The process to determine or extract bullish/bearish sentiment from a tweet is previously discussed in Chapter 2. A measure that is more than 0 is bullish, while 0 is neutral and less than 0 is bearish. This measure accounts for large number of tweets expressing a particular sentiment. The authors found this measure to be most robust out of all indices proposed in their study.

I generated the disagreement index measure (Equation 3) which was introduced by Das and Chen (2007). This measure lies between 0 (no disagreement) and 1 (total

disagreement). This measure helps us to understand more about the relationship between tickers or authors that are highly extreme (very bullish or very bearish) or mixed (equally balanced between bullish and bearish postings) with predictive accuracy. I calculate the bullishness index and disagreement index per period per ticker and generate the average for each individual.

$$\text{Disagreement Index} = \left| 1 - \left| \frac{M^{BULL} - M^{BEAR}}{M^{BULL} + M^{BEAR}} \right| \right| \quad \text{(Equation 3)}$$

Generating Sentiment Similarity Measures

I extend the basic sentiment measures by accounting for individual's sentiment similarity with peers, based on the retweet, mention and reply networks. Specifically, I introduce two new measures: sentiment similarity index (SS) and average sentiment distance (SD). The intuition behind these measures is that through the course of conversations people tend to gravitate towards peers with the same sentiment (Gu et al., 2007). Thus I assert that similarity can further explain influence as well as information quality. I operationalized these measures as follows:

Sentiment Similarity Index (SS)

SS(A) = average similarity (A,*i*) for all *i* that A interacts with. (Equation 4)

where similarity is 1 if author A and peer *i* is of the same bullishness index direction (bullish or bearish) and 0 otherwise.

Average Sentiment Distance (SD)

SD(A) = (SUM (bullish_index(*i*) – bullish_index(A))) / N of *I* (Equation 5)

In essence the average sentiment distance between author A and all his/her peers (*i*).

An example is provided in (See Appendix page 171). A descriptive statistics for all measures is listed in Table 8.

Empirical Methodology and Results

In this section I discuss the models used to test the hypotheses and the corresponding results. I tested three groups of models: OLS, Random Effects and Fixed Effects panel time-sequencing regression models (Allison, 2009) with 10 bootstrap replications in the Stata statistical package to examine the relationships between individual characteristics, peer influence and information quality in the context of stock microblogging. The primary interest is on the RE models because the focus is on inferences about the population rather than individual subjects (Frees, 2004). RE assumes that measures are randomly sampled from a larger population, thus the variances between subjects are interesting and representative of the population. In contrast, the FE models assume that subjects are fixed therefore the differences between subjects are ignored. Even though I discuss RE models in this section, I report all results as well. Included as controls are 1-2 days lagged measures of peer influence and information quality to control for autocorrelation effects that are common in the stock investing related data (Antweiler & Frank, 2004; Das & Chen, 2007; Sprenger & Welpe, 2010), 1-2 days lagged effort measures (RT_out, mention_out and reply_out) to control for effects of endogeneity (Manski, 1993), lagged follower and following, investor sentiment and

sentiment similarity. As all continuous variables in the dataset are highly skewed, a log (natural) transformation is applied to achieve normality of the distribution.

Individual Characteristics with Peer Influence

I first examine the relationship between individual characteristics and peer influence. The dependent variable for this model is $\log(INFLUENCE^{P}_{it})$ which is the log of peer influence measure of type *p* (i.e. RTD, MD, etc.) for individual *i* in time *t*. The following RE model (Equation 6) is estimated:

$$\log(INFLUENCE^{P}_{it}) = \beta_0 + \beta_1 * SELF_{it} + \beta_2 * RECOGNITION_{it} + \beta_3 * EFFORT_{it-1} + \beta_4 * EFFORT_{it-2} + \beta_5 * \log(INFLUENCE^{P}_{it-1}) + \beta_6 * \log(INFLUENCE^{P}_{it-2}) + \beta_7 * CONTROLS_{it} + \mu_i + \epsilon_{it}$$

(Equation 6)

where:

β_0 is the intercept.

$log(INFLUENCE^{P}_{it-1})$ and $log(INFLUENCE^{P}_{it-2})$ are continuous one and two-day lagged variables for peer influence measure of the same type *p*.

$SELF_{it}$, $RECOGNITION_{it}$, $EFFORT_{it-1}$ and $EFFORT_{it-2}$ are vector of variables of the respective type. Specifically $EFFORT_{it-1}$ and $EFFORT_{it-2}$ are one and two-day lagged variables for individual effort measures.

$CONTROLS_{it}$ is a vector of control variables.

μ is an individual fixed effect that controls for the individual differences.

ϵ is the error term.

The primary objective is to measure $\beta_{1,}\beta_{2,}\beta_{3}$ and $\beta_{4,}$ which are the coefficients for individual characteristics (self-disclosed, effort and recognition). The results for estimating this model are in Table 9. In this study I focus on residual values (RTD, MD and ReD) for the Random Effect models as they are most salient in explaining the relationship between peer influence and information quality. Nevertheless, results for other DV are also reported. The values for adjusted R-squared from the results (*RTD*=.6, MD =.6 and ReD=.4) are very encouraging.

For the self-disclosed group, disclosing both *demographics* (RTD .028, MD .051) and *trading preferences* (RTD .029, MD .083) are positively correlated with peer influence. ReD do not show a significant correlation with peer influence. Thus, H1A is supported as presence of self-disclosures positively correlate with peer influence.

For the recognition group, the *suggested* individuals are positively correlated with peer influence. All three residual peer influence values are highly significant with strong correlation values (RTD .528, MD 1.175 and ReD .195). Hence H2A is supported as recognition positively correlates with peer influence.

For individual engaging effort, outdegree-*retweets* (for *t*-1 RTD .068; for *t*-2 RTD .034), *outdegree-mentions* (for *t*-1 *NS; for *t*-2 MD .026), *outdegree-reply* (for *t*-1 ReD .065; for *t*-2 ReD .045) are positively correlated with peer influence. Total tweets are also positively correlated (RTD .14; MD .158; ReD .128). Thus H3A is supported as higher effort measures positively correlate with peer influence.

As expected, lagged indegree *peer influence* measures are all positively correlated with *peer influence* (for *t*-1 RTD .448, MD .157 and ReD .263; for *t*-2 RTD .265, MD .124 and ReD .226). So are sentiment similarity (RTD .121; MD -.143; ReD .177) and

sentiment distance (RTD .159; MD .158; ReD .22). However, lagged followers (for t-1 RTD -.039; MD -.002; ReD -.014) and followings (for t-1 RTD -.007; MD -.013; Red .005) are both generally negatively correlated with peer influence. As for investor sentiment, bullishness index (RTD -.02; MD NS; ReD -.009) is inversely correlated with peer influence. Thus H5A is supported as bearish sentiment correlates with higher peer influence. At the same time, disagreement index (RTD .01; MD .011; ReD.006) is positively correlated signifying that sentiment of influential people tend to spread out and less polarized (See Appendix page 175-179).

Individual Characteristics with Information Quality

Thereafter I examine the relationship between individual characteristics and information quality. The dependent variable for this model is InQ_{it} which is the average information quality measure for individual *i* at time *t*. I focus on p0 and p1 as information quality in the later days are less important to the estimation. The following RE model (Equation 7) is estimated:

$$InQ_{it} = \beta_0 + \beta_1 * SELF_{it} + \beta_2 * RECOGNITION_{it} + \beta_3 * EFFORT_{it\ -1} + \beta_4 * EFFORT_{it-2} + \beta_5 * InQ_{it-1} + \beta_6 * InQ_{it-2} + \beta_7 * CONTROLS_{it} + \mu_i + \epsilon_{it}$$

(Equation 7)

where:

β_0 is the intercept.

InQ_{it-1} and InQ_{it-2} are continuous one and two-day lagged variables for information quality.

$SELF_{it}$, $RECOGNITION_{it}$ and $EFFORT_{it-1}$ and $EFFORT_{it-2}$ are vector of variables of the respective type.

$CONTROLS_{it}$ is a vector of control variables.

μ is an individual fixed effect that controls for the individual differences.

ϵ is the error term.

See Table 10 for results. The primary objective is to measure $\beta_{1,}$ $\beta_{2,}$ β_3 and β_4 the coefficients for the different groups of individual characteristics. The values for adjusted R-squared from the results (p_0 =.11, p_1 =.12) are similar to previous literature examining information quality in stock message board posts (Sprenger & Welpe, 2010 and Antweiler & Frank, 2004). As a consensus, these low values support the notion that information quality is hard to explain. I observe the following in the Random Effect regression results.

For the self-disclosed group, disclosing of *demographic* is not significant whereas disclosing of *trading preferences* is positively correlated with *information quality* (p0 .012; p1 .016). In essence those who disclose trading preferences are more likely to possess higher information quality as compared to demographic preferences. Trading preferences seem to have more information that relate to higher information quality. Thus H1B is supported.

For the recognition group, being an expert (*suggested*) (*p0* -.04, *p1* NS) is negatively correlated with *information quality*. This is interesting as intuition tells us that those who are labeled as experts by authority should possess higher information quality but it turned out to be the reverse. Thus H2B is supported.

For the engaging effort group, both *reply-out* (*t-2 p1* -.008) and *RT-out* (*t*-1 *p0* -.009; *p1* -.007) are weakly negatively correlated with information quality while mention is not significant. However, total tweets are positively correlated with information quality (RE p0 .05; p1 .101). Thus H3B is supported.

For investor sentiment, both *bullishness index* (*p0* -.002, *p1*-.034) and disagreement index (*p0* -.295; *p1* -.301) have negative correlations with information quality. H5B is supported. Contrary, *lagged information quality* measures are all positively significant confirming the impact of auto-correlation. Parallel, lagged follower and following are generally inversely correlated while sentiment distance is not significant with information quality (See Appendix page 181).

Peer Influence and Information Quality

I finally examine the relationship between individual characteristics and peer influence, and extending to information quality, the main focus of this study. The dependent variable is InQ_{it} which is the average information quality measure for individual *i* at time *t*. The following RE model (Equation 8) is estimated:

$$InQ_{it} = \beta_0 + \beta_1 * SELF_{it} + \beta_2 * RECOGNITION_{it} + \beta_3 * EFFORT_{it\ -1} + \beta_4 * EFFORT_{it-2} + \beta_{5*} InQ_{it-1} + \beta_6 * InQ_{it-2} + \beta_7 * log(INFLUENCE^{P}_{it-1}) + \beta_8 * log(INFLUENCE^{P}_{it-2}) + \beta_9 * CONTROLS_{it} + \mu_i + \epsilon_{it} \quad \text{(Equation 8)}$$

where:

β_0 is the intercept.

InQ_{it-1} and InQ_{it-2} are continuous lagged variables for information quality while $log(INFLUENCE^{P}_{it-1})$ and $log(INFLUENCE^{P}_{it-2})$ are continuous one and two-day lagged variables for peer influence measure of the same type p.

$SELF_{it}$, $RECOGNITION_{it}$ and $EFFORT_{it-1}$ and $EFFORT_{it-2}$ are vector of variables of the respective type.

$CONTROLS_{it}$ is a vector of control variables.

μ is an individual fixed effect that controls for the individual differences.

ϵ is the error term.

The primary objective is to measure β_7 and β_8 which are the coefficients for lagged variables of peer influence of type p. Table 11 presents the results for estimating this model. Adjusted R-squared is comparable with previous literature examining similar information quality (Antweiler & Frank, 2004; Sprenger & Welpe, 2010) (Adjusted R-squared *p0* =.12, *p1* =.13).

Results are salient and interesting. Lagged indegree residual peer influence measures for RTD and MD are significant, but ReD is not significant. Lagged RT (p0 -.019) is negatively correlated with information quality while MD (p0 .014) is positively correlated with information quality. Relationships of different types of peer influence with information quality differ. Thus, both H4A and H4B are supported. For normalized influence by peer outdegree and normalized influence by unique peers, however, lagged reply peer influence is significant (p0 .028 and .032). Thus when peers' attention is accounted for, reply peer influence is more salient than mention and retweets. Table 12 lists the results for the hypotheses summary (See Appendix page 182 and 183 for additional results).

Discussion

Individual Characteristics and Peer Influence

The results confirm that both self-disclosed individual demographic attributes and trading preferences are highly significant in explaining peer influence. This supports social identity theory (Ellemers et al., 1999; Tajfel, 1978) which states that individuals affirm a clear and consistent sense of self and wish to feel connected to others in receiving identity-affirming feedback (Forman et al., 2008). Such self-disclosed information help peers to identify and connect with each other through demographic information such as name, location, bio or url or via trading preferences such as investing portfolio, strategies and approaches. By reaching a consensus regarding identities, people feel understood and obtain a sense of continuity and coherence (Ma & Agarwal, 2007; Swann et al., 2000), facilitate formation of relationships, common bonds and social attractions that community members value (Ren et al., 2007). In short, the desire to maintain connectivity through disclosing both demographic and trading preferences is a strong motivation that leads to peer influence.

In terms of recognition, firm recognized “suggested” label is deemed to be influential and is consistently recognized as such. In Stocktwits where information overload is a norm, people consistently seek out parsimonious signals of influence and information quality. Since the suggested label is assigned by Stocktwits, which is an authoritative figure in the eyes of the community, those with this “firm recognized” status are perceived to be trustworthy and influential (Jeppesen & Frederiksen, 2006). This is in line with Forman et al. (2008) which state that individuals who are recognized for their expertise possess high levels of perceived credibility. Thus heuristic information

processing (Chaiken & Maheswaran, 1994), when people formed or changed their attitudes by invoking heuristics such as “those in the suggested list are trustworthy”, plays a major role. In the Stocktwit community, where cognitive capacity is a constraint, peers resort to such mental shortcuts in determining credible sources.

In terms of individual effort, I note that those who consistently invest effort to engage and contribute knowledge are deemed influential. Due to reciprocity (Gouldner, 1960),) the effort of these individuals are acknowledged and appreciated. Such positive relationship with peer influence is not gained spontaneously or accidently but through concerted effort (Chai et al., 2012) and consistent personal involvement. This is in line with social exchange theory (Emerson, 1976) which states that people participate expecting social rewards and it turn do receive approval, status and respect from peers.

Individual Characteristics and Information Quality

The results conclude that it is very challenging to explain information quality. This is probably due to high uncertainty derived from the difficulty and subjectivity in determining value of stocks (Baker & Wurgler, 2007).

In terms of self-disclosed individual attributes, I found that the disclosure of trading preferences is salient towards explaining information quality while disclosing of demographic information is not. Thus it is not only that one has to reveal preferences but those preferences should be relevant to the context in question. This is related to the source credibility theory (Chaiken & Maheswaran, 1994) where those who reveal more personal identifiable information have higher credibility (Forman et al, 2008). Individuals who disclose their trading experiences are more committed and dedicated to the domain

of stock investing, thus relate to higher information quality. As with Forman et al. (2008) who found that reviews with identity-descriptive information are read more positively and associated with subsequent increase in product sales, I conclude that tweets from identifiable trading preferences contain higher information quality. Furthermore, the level of accountability for these individuals may also increase due to copresence (Golfman, 1959) especially within the small-knit community such as Stocktwit. In short, just like self-disclosure's relationship with peer influence, the desire to maintain connectivity through disclosing trading preferences is a strong motivation that leads to information quality.

In terms of recognition attributes, suggested is negatively correlated to information quality. Seemingly expert individuals are not the ones with information quality. The information processing theory (Chaiken & Maheswaran, 1994) has clearly shown that when peers rely on heuristic processing to determine credible sources, their ability to identify sources with information quality deteriorates. Furthermore, experts are determined by Stocktwits which has its own business rules in selecting individuals as experts. These rules are in a black box that may not be aligned with how investors in the community evaluate peers. In addition, being assigned as an expert persists over time but information quality is real-time and thus fluctuates with time and market dynamics. So an individual who was active and with high information quality before but is now less active may still be on the suggested list but now has low information quality.

In terms of individual effort, engaging with others (outdegree retweets, mentions and replies) is a negative predictor of information quality which tells us that there is more than just exchange of trading information in these social interactions. Social

psychological phenomena from behavioral finance may play a role in impacting the quality of trading information in the community. Investor overconfidence (Hirshleifer, 2001) leads to high engaging effort that unfortunately correlates with low information quality. Perhaps these activities focus on the need to connect with peers but neglect information quality of the content being shared.

Peer Influence and Information Quality

The relationship between peer influence and information quality is salient but differs according to the measures being used. Two conclusions are determined. First, individuals who received higher residual indegree retweets are likely to associate with low information quality. Second, individuals with high indegree mentions are those with high information quality. In short, those whose tweets to his or her followers that are retweeted often are likely to possess low information quality while those mentioned frequently by peers are likely to possess higher information quality. Clearly these are two different peer influence dimensions and should not be generalized as one. Retweet is due to the individual's pass along or diffusion ability (Zhu & Chau, 2012) while mention is due to the individual's ability to get noticed associated with certain expertise or other intrinsic values that the individual may possess.

The explanation is both intriguing and interesting. Retweet is driven by content value of the tweet while mention is recognition on an individual or name value of the individual (Cha et al., 2010; Leavitt et al., 2009). Online investors face high information overload (Brynjolfsson & Smith, 2000) and has limited attention and processing power/cognitive capacity (Hirshleifer & Teoh, 2003; Metzger et al., 2010). Due to such

constraints, people are less likely to remember tweet over time but are more likely to remember those individuals who consistently share and publish quality trading information, as per source credibility theory (Hass, 1981). Thus mention is a good predictor of peers with information quality. The fact is that influential people do have information quality and the best way to identify them is to let the community point out who they are through mentions. Retweet influence, on the other hand, is highly swayed by an individual's network susceptibility (Watts & Dodds, 2007) as well as his personal traits. Obviously, he or she has the ability to diffuse information to his/her followers at a higher rate than others. But since peers' followers are more likely to be similar in preferences and beliefs due to homophily (Chen et al., 2009), high retweets relate to low information quality. Overall, it is intriguing that microblogging measures are able to discern the effects of peer influence at such a low level of granularity.

Contributions and Implications

This study's first contribution is to explain the relationship between peer influence and information quality in the context of individual characteristics of an online investing community. Surprisingly, the results show that different facets of investor characteristics have a different relationship with information quality. Specifically the set of individual characteristics that relate to peer influence is not synonymous with the set of individual characteristics that explain information quality. Furthermore, peer influence measures for source (mentions) are more salient in explaining higher information quality as compared to those measures for content (retweets). This study presents a new dimension for scholars involved in research on peer influence and information quality.

Current peer influence research ignores information quality perhaps due to the challenge in measuring information quality. As the second contribution in this study, I demonstrate an intrinsic measure of information quality as an initial effort in extending peer influence with information quality. I urge scholars to expand the current work by investigating other types of information quality such as contextual, representational and accessibility (Wang & Strong, 1996). Clearly, information quality plays a major role in the peer influence model.

As the third contribution, this study presents an explanation of the relationship between peer influence and information quality that helps to guide VIC investors to identify high information quality peers in the community. People participate in online communities with different motivations. People interact not just for information exchange but also for social support (Wasko & Faraj, 2005). Thus, those seeking social needs should consider peers who are influential: on the suggested list and continuously engaging in sending retweets, mentions and replies. Conversely those seeking higher information quality should consider peers who are mentioned frequently by others and those posting a higher volume of stock tweets. Consistent with findings of Cha et al. (2010), those who consistently invest effort in sharing information are more likely to be associated with both high peer influence and high information quality.

For VIC managers I tested a better measure of peer influence, namely the residual measures, for identifying peer influence as well as the type of author characteristics that relate to information quality and peer influence as the fourth contribution of this study. These measures were validated to be reliable and able to explain both peer influence and information quality. These measures may aid managers in designing new features to

enhance the usefulness of the communities and providing valuable investing information for investors. In addition, this may lead to monetization incentives in packaging tweets from those with higher influence or information quality. Gaining this understanding is important because it guides manager on how to make virtual communities relevant/useful and influential for their participants.

Limitations and Future Research

This study, like others, does not come without limitations. First of all, the measure of intrinsic information quality (Wang & Strong, 1996) in this study is highly objective and convenient but alludes to a very narrow scope of information quality, specifically covering only tweets with sentiment. Some tweets in the sample are neutral but they may contain interesting insights or include links to rich information such as charts and useful investing information. In the current definition of information quality, such tweets are not considered. Essentially quality investing information should not be limited to sentiment only, but should include information that helps investors to gain knowledge in investing. In future work, I might consider contrasting information quality of different tiers or examine information quality of a wider scope.

Second, an important extension of this study is to include news, analyst forecasts and stock information as controls. It is interesting to understand how these external sources impact peer influence and information quality in the community. I could also include different characteristics of stocks (e.g., small and large capitalization, etc.) as well as other investing instruments such as foreign exchange and futures to further enrich

the dataset and discover deeper relationships between influence and information quality with stock characteristics.

Another pertinent extension is to explore network analysis within the community relating to peer influence and information quality. From the current study I know that attributes of different authors have different impact on their influence. But I have only scratched the surface in terms of using the social measures of retweets, mentions and replies. I can examine connections between actors in the social network (Grannovetter, 1992) to a few layers deep to extract hidden measures that may better explain peer influence and information quality.

Conclusion

Many have investigated the notion of peer influence but few examined the relationship between peer influence and information quality. I study this relationship in the context of individual characteristics in stock microblogging. Surprisingly, I discover that the set of individual characteristics that relate to peer influence is not synonymous with those that relate to high information quality. Specifically, to be perceived as influential one has to be willing to disclose both demographic and trading preferences, be on the suggested list and to have close sentiment similarity with peers. On the contrary, disclosing trading preferences correlate with high information quality, being on the suggested list implies low information quality while sentiment similarity is not relevant. In comparing peer influence and information quality, however, I establish that those who are frequently mentioned by peers are likely to possess higher information quality while those whose tweets are frequently retweeted are associated with low information quality.

Furthermore, this study contains three important contributions. First, I describe the relationships between individual characteristics, peer influence and information quality and in so doing extend the peer influence model with information quality, an IS research extension that is largely unexplored. Second, this study is among the first to examine peer influence in the financial stock investing domain of stock microblogging, and in identifying different predictors for influencers and high information quality individuals from those in other contexts. In so doing, it contributes to the behavioral finance literature by providing empirical evidence of the impact of investor sentiment (Barberis et al., 1998). Third, I present practical measures of peer influence to researchers and practitioners such as investors and VIC managers, which measures that are more relevant to platform or individual decisions in microblogging-enabled VICs.

Table 5. Hypotheses summary

	Hypothesis	Related Theory
H1A	All else equals, demographic and trading preferences self-disclosures should positively correlate with peer influence.	Social identify
H1B	All else equals, demographic and trading preferences self-disclosures should positively correlate with information quality.	Source credibility and copresence
H2A	All else equals, being on the suggested list should positively correlate with peer influence.	Information processing theory (systematic and heuristic theory)
H2B	All else equals, being on the suggested list should negatively correlate with information quality.	Information processing theory (systematic and heuristic theory)
H3A	All else equals, measures of effort should positively correlate with peer influence.	Social exchange theory
H3B	All else equals, measures of effort should negatively correlate with information quality.	Illusion of control Investor overconfidence
H4A	All else equals, individuals with higher peer influence measure of indegree mentions should be positively correlated with information quality.	Source credibility
H4B	All else equals, individuals with higher peer influence measure of indegree retweets should be negatively correlated with information quality.	Homophily and similarity
H5A	All else equals, investor sentiment should be negatively correlated with peer influence.	Prospect theory
H5B	All else equals, investor sentiment should be negatively correlated with information quality.	Wishful thinking and short-selling

Table 6. Dependent variables

	Abbre.	DV Type	Description
Group 1	RTI	RT_in	Baseline indegree count of retweets
	MI	Mention_in	Baseline indegree count of mentions
	ReI	Reply_in	Baseline indegree count of replies
Group 2	RTD	RT_diff	Residual of indegree and outdegree retweets
	MD	Mention_diff	Residual of indegree and outdegree mentions
	ReD	Reply_diff	Residual of indegree and outdegree replies
Group 3	RTN	RT_normalized	Normalized residual retweets by total tweets
	MN	Mention_normalized	Normalized residual mentions by total tweets
	ReN	Reply_normalized	Normalized residual replies by total tweets
Group 4	RT_NIPO	RT_norm_influence_by_ peer_outdegree	Normalized of indegree retweets by peers' outdegree
	M_NIPO	Mention_norm_influence_by_ peer_outdegree	Normalized of indegree mentions by peers' outdegree
	Re_NIPO	Reply_norm_influence_by_ peer_outdegree	Normalized of indegree replies by peers' outdegree
Group 5	RT_NIUP	RT_norm_influence_by_ unique_peers	Normalized of indegree author count by peers' outdegree author count (retweet network)
	M_NIUP	Mention_norm_influence_by_ unique_peers	Normalized of indegree author count by peers' outdegree author count (mention network)
	Re_NIUP	Reply_norm_influence_by_ unique_peers	Normalized of indegree author count by peers' outdegree author count (reply network)

Table 7. Independent variables

	Abbre.	DV Type	Description
Self-disclosed	DEMO	demographic_disclose	Self-disclosure of demographic information.
	TRAD	trading_disclose	Self-disclosure of trading information
Recognition	SUG	suggested	Being on the suggested list
Effort	RTO	RT_out	Outdegree retweets
	MO	mention_out	Outdegree mentions
	ReO	reply_out	Outdegree replies
	TOT	total_tweets	Total tweets sent
Controls	BULL	bullish_index	Investor sentiment of bullishness
	DIS	disagree_index	Polarity of sentiment
	FO	follower	Count of followers
	FI	following	Count of followings
	RT_SS	RT_sentiment_similariy	Sentiment similarity by Retweets
	M_SS	mention_sentiment_similariy	Sentiment similarity by Mentions
	Re_SS	reply_sentiment_similariy	Sentiment similarity by Replies
	RT_SD	RT_sentiment_distance	Sentiment distance by Retweets
	M_SD	mention_sentiment_distance	Sentiment distance by Mentions
	Re_SD	reply_sentiment_distance	Sentiment distance by Replies
	None	Lagged peer influence DV	Lagged peer influence DV
	None	Lagged information quality DV	Lagged information quality DV

Table 8. Descriptive statistics

		Abbr.	N	Minimum	Maximum	Mean	Std. Deviation
Dependent Variables							
Peer Influence DV	RT_in	RTI	47973	0	126	1.40	5.237
	mention_in	MI	47973	0	342	4.62	15.808
	reply_in	ReI	47973	0	99	.70	2.361
	RT_diff	RTD	47973	-152	97	.57	4.462
	mention_diff	MD	47973	-182	261	3.48	12.955
	reply_diff	ReD	47973	-157	93	-.16	2.741
	RT_norm	RTN	47973	-1.0000	97.0000	.122363	.8964139
	mention_norm	MN	47973	-1.0000	149.0000	.564238	2.3221140
	reply_norm	ReN	47973	-1	15	-.02	.364
	RT_norm_influence_peer_outdegree	RT_NIPO	47973	0.000000	1.000000	.06061017	.184826215
	mention_norm_influence_peer_outdegree	M_NIPO	47973	0.000000	1.000000	.06530189	.194830406
	reply_norm_influence_peer_outdegree	Re_NIPO	47973	0	1	.09	.226
information quality DV	p0	p0	47973	0.000000	1.000000	.35315929	.379042576
	p1	p1	47973	0.000000	1.000000	.37721179	.387546425
Independent Variables							
Self-disclosed	demographic_disclose	DEMO	47973	0	1	.87	.332
	trading_disclose	TRAD	47973	0	1	.84	.363
recognition	suggested	SUG	47973	0	1	.06	.239
Effort	RT_out	RTO	47973	0	234	.83	4.405
	mention_out	MO	47973	0	234	1.13	5.654
	reply_out	ReO	47973	0	184	.86	3.276
	total_tweets	TOT	47973	1	1302	10.17	25.933
	bullish_index	BULL	47973	-3.8501	4.8598	.535192	.7265822
	disagree_index	DIS	47973	0.0000	1.0000	.396705	.4184198
Controls	follower	FO	47973	0	32073	153.15	991.543
	following	FI	47973	0	11063	46.38	235.385
	RT_sentiment_similariy	RT_SS					
	mention_sentiment_similariy	M_SS	47973	0	1	.12	.320
	reply_sentiment_similariy	Re_SS	47973	0	1	.14	.340
	RT_sentiment_distance	RT_SD	47973	0	4	.12	.354
	mention_sentiment_distance	M_SD	47973	0	5	.14	.384
	reply_sentiment_distance	Re_SD	47973	0	5	.18	.430
	Lagged peer influence DV	As per each peer influence DV					
	Lagged information quality DV	As per each information quality DV					

Table 9. OLS, Random Effect and Fixed Effect results for relating individual characteristics with peer influence.

		OLS			Random Effects			Fixed Effects		
		RTD	MD	ReD	RTD	MD	ReD	RTD	MD	ReD
Self-disclosed	DEMO	0.028 (0.006) ****	0.056 (0.008) ****	-0.005 (0.005)	0.028 (0.006) ****	0.051 (0.009) ****	-0.005 (0.005)			
	TRAD	0.029 (0.005) ****	0.069 (0.007) ****	-0.002 (0.004)	0.029 (0.005) ****	0.083 (0.009) ****	-0.002 (0.004)			
Recognition	SUG	0.528 (0.021) ****	0.762 (0.023) ****	0.195 (0.014) ****	0.528 (0.021) ****	1.175 (0.038) ****	0.195 (0.014) ****	0.117 (0.069) *	0.1 (0.037) ***	0.035 (0.071)
Effort	Lag t-1DV out degree	0.068 (0.017) ****	0.012 (0.01)	0.065 (0.018) ****	0.068 (0.017) ****	-0.001 (0.009)	0.065 (0.018) ****	0.003 (0.016)	-0.034 (0.007) ****	0.058 (0.019) ***
	Lag t-2 DV out degree	0.034 (0.016) **	0.042 (0.009) ****	0.045 (0.018) **	0.034 (0.016) **	0.026 (0.009) ***	0.045 (0.018) **	-0.019 (0.016)	-0.012 (0.007) *	0.039 (0.02) **
	TOT	0.14 (0.004) ****	0.158 (0.005) ****	0.128 (0.004) ****	0.14 (0.004) ****	0.142 (0.004) ****	0.128 (0.004) ****	0.121 (0.005) ****	0.055 (0.004) ****	0.163 (0.005) ****
Controls	BULL	-0.02 (0.004) ****	-0.015 (0.005) ****	-0.009 (0.003) ***	-0.02 (0.004) ****	-0.003 (0.004)	-0.009 (0.003) ***	-0.015 (0.004) ****	0.007 (0.003) ***	-0.011 (0.004) ***
	DIS	0.01 (0.005) *	0.037 (0.007) ****	0.006 (0.004) *	0.01 (0.005) *	0.011 (0.005) **	0.006 (0.004) *	-0.012 (0.005) **	-0.007 (0.004) *	0.005 (0.005)
	FO1	-0.039 (0.002) ****	-0.062 (0.003) ****	-0.014 (0.002) ****	-0.039 (0.002) ****	-0.02 (0.002) ****	-0.014 (0.002) ****	-0.026 (0.003) ****	0.007 (0.002) ****	-0.001 (0.002)
	FO2	-0.009 (0.003) ***	-0.037 (0.003) ****	0.001 (0.003)	-0.009 (0.003) ***	-0.011 (0.003) ****	0.001 (0.003)	-0.008 (0.004) **	0.006 (0.002) **	0.002 (0.003)
	FI1	-0.007 (0.001) ****	-0.013 (0.002) ****	0.005 (0.001) ****	-0.007 (0.001) ****	-0.013 (0.002) ****	0.005 (0.001) ****	-0.004 (0.002) ***	0.001 (0.001)	0.002 (0.001)
	FI2	-0.002 (0.002)	-0.009 (0.003) ****	-0.017 (0.002) ****	-0.002 (0.002)	-0.003 (0.003)	-0.017 (0.002) ****	0.012 (0.003) ****	-0.002 (0.002)	-0.002 (0.003)
	Lag t-1DV in-degree	0.448 (0.011) ****	0.467 (0.01) ****	0.263 (0.012) ****	0.448 (0.011) ****	0.157 (0.008) ****	0.263 (0.012) ****	0.159 (0.012) ****	-0.012 (0.005) **	0.012 (0.012)
	Lag t-2 DV in degree	0.265 (0.011) ****	0.376 (0.01) ****	0.226 (0.012) ****	0.265 (0.011) ****	0.124 (0.008) ****	0.226 (0.012) ****	0.014 (0.011)	-0.003 (0.005)	0.007 (0.013)
	SS	0.121 (0.034) ****	0.031 (0.026)	0.177 (0.029) ****	0.121 (0.034) ****	-0.143 (0.02) ****	0.177 (0.029) ****	-0.098 (0.029) ****	-0.235 (0.015) ****	0.128 (0.031) ****
	SD	0.159 (0.035) ****	0.067 (0.026) ***	0.22 (0.024) ****	0.159 (0.035) ****	-0.158 (0.019) ****	0.22 (0.024) ****	-0.201 (0.028) ****	-0.29 (0.014) ****	0.087 (0.027) ****
	_cons	-0.046 (0.008) ****	0.056 (0.011) ****	-0.076 (0.007) ****	-0.046 (0.008) ****	0.085 (0.011) ****	-0.076 (0.007) ****	0.171 (0.009) ****	0.725 (0.006) ****	-0.095 (0.009) ****
	R2	0.6	0.73	0.4	0.6	0.6	0.4	0.43	0.009	0.29
	N	38865	40947	34302	38865	40947	34302	38865	40947	34302

Table 10. OLS, Random Effect and Fixed Effect regression results for relating individual characteristics with information quality

		OLS		RE		FE	
	DV	p0	p1	p0	p1	p0	p1
Self-Disclosed	DEMO	0 (0.005)	0.005 (0.005)	-0.004 (0.007)	0.001 (0.007)		
	TRAD	0.014 (0.005) ***	0.018 (0.005) ****	0.012 (0.006) *	0.016 (0.007) **		
Recognition	SUG	-0.021 (0.007) ***	-0.008 (0.007)	-0.04 (0.012) ****	-0.016 (0.012)	0.004 (0.038)	-0.028 (0.037)
Effort	RTO1	0.006 (0.007)	0.014 (0.007) *	0 (0.007)	0.007 (0.008)	-0.002 (0.009)	0.002 (0.009)
	RTO2	-0.005 (0.007)	-0.011 (0.008)	-0.01 (0.008)	-0.016 (0.008) **	-0.013 (0.009)	-0.019 (0.009) **
	MO1	-0.009 (0.006)	-0.016 (0.007) **	-0.003 (0.007)	-0.009 (0.007)	0.002 (0.008)	-0.003 (0.008)
	MO2	-0.004 (0.007)	0.003 (0.007)	0.002 (0.007)	0.008 (0.007)	0.005 (0.008)	0.012 (0.008)
	ReO1	-0.009 (0.004) **	-0.007 (0.004) *	-0.009 (0.004) **	-0.007 (0.004) *	-0.007 (0.004) *	-0.006 (0.004)
	ReO2	-0.003 (0.004)	-0.003 (0.004)	-0.003 (0.004)	-0.003 (0.004)	-0.003 (0.004)	-0.003 (0.005)
	TOT	0.098 (0.002) ****	0.096 (0.002) ****	0.105 (0.002) ****	0.101 (0.002) ****	0.102 (0.003) ****	0.098 (0.003) ****
Controls	BULL	-0.062 (0.003) ****	-0.036 (0.003) ****	-0.062 (0.003) ****	-0.034 (0.003) ****	-0.054 (0.004) ****	-0.026 (0.004) ****
	DIS	-0.303 (0.005) ****	-0.31 (0.005) ****	-0.295 (0.005) ****	-0.301 (0.005) ****	-0.261 (0.006) ****	-0.271 (0.006) ****
	FO1	-0.001 (0.002)	-0.002 (0.002)	0.003 (0.002) **	-0 (0.002)	0.01 (0.002) ****	0.001 (0.002)
	FO2	-0.007 (0.002) ****	-0.007 (0.002) ****	-0.006 (0.002) ***	-0.006 (0.002) ***	-0.004 (0.003)	-0.004 (0.003)
	FI1	-0.001 (0.001)	-0.001 (0.001)	-0.003 (0.001) **	-0.001 (0.001)	-0.005 (0.002) ***	0.001 (0.002)
	FI2	0.005 (0.002) ***	0.007 (0.002) ****	0.009 (0.002) ****	0.007 (0.002) ****	0.013 (0.003) ****	0.007 (0.003) **
	Lag t-1 DV	0.036 (0.006) ****	0.043 (0.006) ****	-0.018 (0.006) ***	0.007 (0.006)	-0.086 (0.007) ****	-0.031 (0.007) ****
	Lab t-2 DV	0.024 (0.006) ****	0.002 (0.006)	-0.028 (0.006) ****	-0.036 (0.006) ****	-0.085 (0.007) ****	-0.07 (0.007) ****
	RT_S	0.003 (0.012)	0.014 (0.013)	-0.005 (0.012)	0.005 (0.013)	-0.002 (0.015)	0.001 (0.015)
	M_S	-0.008 (0.011)	-0.032 (0.012) ***	-0.014 (0.011)	-0.036 (0.012) ***	-0.023 (0.013) *	-0.04 (0.013) ***
	Re_S	0.006 (0.006)	-0.001 (0.006)	-0.003 (0.006)	-0.007 (0.007)	-0.005 (0.008)	-0.006 (0.008)
	_cons	0.334 (0.007) ****	0.344 (0.007) ****	0.336 (0.009) ****	0.342 (0.009) ****	0.339 (0.007) ****	0.357 (0.007) ****
	R2	0.11	0.12	0.11	0.12	0.09	0.11
	N	47973	47973	47973	47973	47973	47973

Table 11. Random Effect regression results for relating individual characteristics and peer influence with information quality

	Influence=t	INDEGREE	RESIDUAL	NORMALIZED	NORMALIZED INFLUENCE by PEER OUTDEGREE	NORMALIZED INFLUENCE by UNIQUE PEER
	DV	p0	p0	p0	p0	p0
Self-disclosed	DEMO	-0.004 (0.007)	0.004 (0.008)	-0.001 (0.008)	-0.004 (0.007)	-0.004 (0.007)
	TRAD	0.012 (0.006) *	0.005 (0.008)	0.009 (0.007)	0.012 (0.006) *	0.012 (0.006) *
Recognition	SUG	-0.041 (0.012) ****	-0.048 (0.014) ****	-0.041 (0.013) ****	-0.041 (0.012) ****	-0.041 (0.012) ****
Effort	TOT	0.107 (0.002) ****	0.107 (0.003) ****	0.098 (0.002) ****	0.105 (0.002) ****	0.105 (0.002) ****
	RTO1	0.001 (0.007)	-0.005 (0.011)	0.003 (0.007)	0 (0.007)	-0 (0.007)
	RTO2	-0.009 (0.008)	-0.015 (0.011)	-0.013 (0.008) *	-0.01 (0.008)	-0.01 (0.008)
	MO1	-0.004 (0.007)	0.003 (0.01)	-0.004 (0.007)	-0.002 (0.007)	-0.002 (0.007)
	MO2	0.002 (0.007)	0 (0.01)	0.005 (0.007)	0.002 (0.007)	0.002 (0.007)
	ReO1	-0.009 (0.004) **	-0.004 (0.006)	-0.008 (0.004) **	-0.009 (0.004) **	-0.009 (0.004) **
	ReO2	-0.004 (0.004)	-0.013 (0.006) **	-0.005 (0.004)	-0.004 (0.004)	-0.004 (0.004)
Controls	BULL	-0.062 (0.003) ****	-0.059 (0.004) ****	-0.058 (0.003) ****	-0.062 (0.003) ****	-0.062 (0.003) ****
	DIS	-0.296 (0.005) ****	-0.298 (0.005) ****	-0.273 (0.005) ****	-0.295 (0.005) ****	-0.295 (0.005) ****
	FO1	0.003 (0.002) *	0.003 (0.002)	0.004 (0.002) **	0.003 (0.002) **	0.003 (0.002) **
	FO2	-0.006 (0.002) ***	-0.007 (0.002) ***	-0.005 (0.002) **	-0.006 (0.002) ***	-0.006 (0.002) ***
	FI1	-0.003 (0.001) **	-0.002 (0.002)	-0.002 (0.001) *	-0.003 (0.001) **	-0.003 (0.001) **
	FI2	0.009 (0.002) ****	0.01 (0.002) ****	0.008 (0.002) ****	0.009 (0.002) ****	0.009 (0.002) ****
	Lagged 1 day DV	-0.018 (0.006) ***	-0.019 (0.007) ***	-0.025 (0.006) ****	-0.018 (0.006) ***	-0.018 (0.006) ***
	Lagged 2 day DV	-0.028 (0.006) ****	-0.032 (0.008) ****	-0.036 (0.006) ****	-0.028 (0.006) ****	-0.028 (0.006) ****
	RT_S	-0.005 (0.012)	0.013 (0.02)	0.005 (0.013)	-0.005 (0.012)	-0.005 (0.012)
	M_S	-0.015 (0.012)	-0.028 (0.018)	-0.017 (0.012)	-0.014 (0.011)	-0.015 (0.011)
	Re_S	-0.003 (0.007)	-0.016 (0.012)	0.002 (0.007)	-0.006 (0.007)	-0.006 (0.007)
Peer influence	RT influence	-0.019 (0.003) ****	-0.019 (0.005) ****	-0.01 (0.009)	-0.002 (0.019)	0.001 (0.019)
	M influence	0.013 (0.004) ****	0.014 (0.005) ***	0.004 (0.007)	-0.02 (0.018)	-0.021 (0.018)
	Re influence	-0.005 (0.004)	0.002 (0.005)	-0.014 (0.009)	0.028 (0.01) ***	0.032 (0.01) ***
	_cons	0.334 (0.009) ****	0.335 (0.011) ****	0.341 (0.01) ****	0.336 (0.009) ****	0.336 (0.009) ****
	R-squared	0.11	0.1	0.09	0.11	0.11
	N	47973	33763	45034	47973	47973

Table 12. Hypotheses results summary

	Peer Influence	**Information Quality**
Self-disclosed	H1A –supported	H1B – supported
Recognition	H2A – supported	H2B – supported
Effort (including sentiment)	H3A – supported H5A – supported	H3B – inconclusive H5B - supported
Peer Influence	NA	H4A – supported H4B –supported

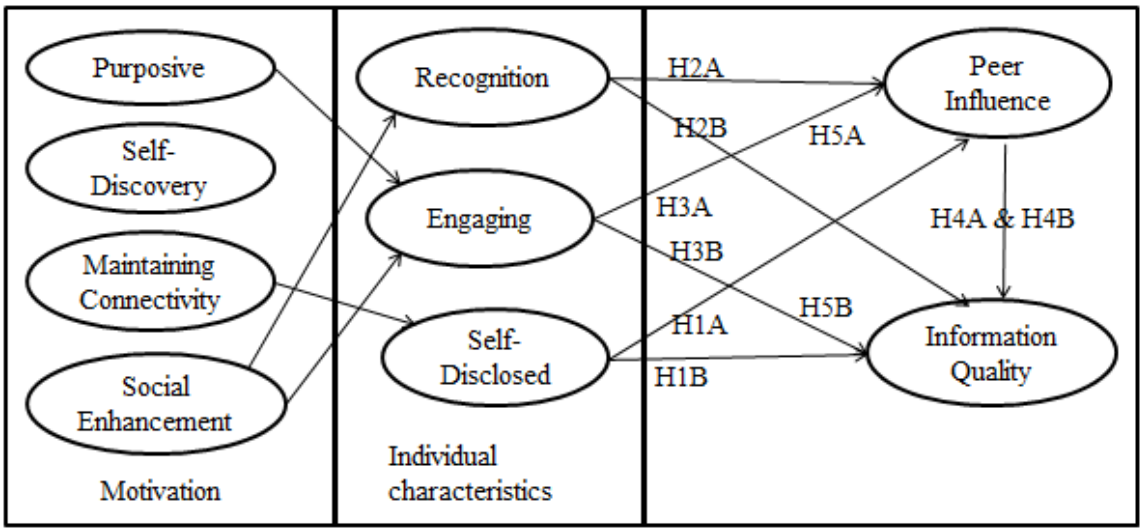


Figure 6. Conceptual framework

CHAPTER 4

INVESTIGATING PREDICTIVE POWER OF STOCK MICROBLOG FEATURES IN FORECASTING FUTURE STOCK PRICE MOVEMENTS

Introduction

The Internet and Stock Investing

The Internet, as a whole, has become an enabler that aggregates vital information for stock investor decision making. It is changing how information is delivered to investors and the ways in which investors can act upon that information (Barber & Odean, 2001). In essence, it alters the way that investors, trade, acquire and share information (Zhang & Swanson, 2010). Initially, it was more a tool for aggregating public information such as financial data, market updates and public news. More recently, with the advent of WEB 2.0 and social media (Baltzan, 2011), user generated content (UGC) are incorporating private information in addition to public information (Tumarkin & Whitelaw, 2001). Thus I observe how virtual investing communities (VIC) Yahoo Finance and Raging Bull are publishing relevant and valuable UGC data such as

investment recommendations and proprietary analysis which enriches investors' ability to make better investment decisions. This allows investors to view the thought process and decision makings of others. As a result, it is imperative for researchers to understand how individuals in virtual communities interact with one another and how these behaviors relate to predictive outcomes in stock performance.

Emergence of Stock Microblogging

With the emergent and popularity of microblogging channels such as Twitter, more and more investors are jumping on the bandwagon in joining their peers in microblogging about stocks real-time, resulting in investing-focused channels such as Stocktwits.com. Microblog's three distinct characteristics - succinctness, high volume and real-time information - greatly facilitate the diffusion of investing information (Bollen et al., 2011; Java et al., 2007). Since microblogs are limited to 140 characters long, they are brief and succinct by design, thus reducing noise and transmitting more relevant messages to their recipients. Additionally, this decrease of time leads to high volume of postings by increasing the frequency of updating a typical microblog, from one a day, as in regular blogging, to multiple per day, in microblogging (Java et al., 2007). Furthermore, postings are real-time as they are posted very close to the occurrence of the events. Being real-time is a key factor in microblogging's popularity (Claburn, 2009) because over time information becomes less relevant and less useful for decision and planning purposes (Ballou & Pazer, 1995). Furthermore, since public investing information such as company press releases, earning announcements and analyst recommendations is sporadic and infrequent, microblogs represent a new source of

information at a constant and greater temporal frequency to the investors. Hence, this study postulates that the microblogging characteristics of succinctness, high volume and real-time may positively contribute to the predictive power of microblog features.

The Efficient Market and Irrational Investor Debate

Does stock investing UGC matter? There are two contrasting views to this question. On one hand, some scholars claim that markets are efficient and security prices always fully reflect all available information (Fama, 1970). Thus 'prices are right' and are set by agents who understand Bayes law and have sensible preferences (Barberis & Thaler, 2003). Specifically, beliefs and behaviors of investors affect others only through market prices, hence they do not disseminate through proximity, whether it be geographically, socially, or connectively via social media (Hirshleifer & Teoh, 2008). On the other hand, behavioral finance scholars postulate that investor beliefs and behaviors do matter and that investor correlated opinions create risks that are associated with market dynamics.

There are two types of traders in the financial markets, the irrational noise traders or day traders (Koski et al., 2004) who hold random beliefs about future dividends--and the rational arbitrageurs who hold Bayesian beliefs (De Long et al., 1990). Noise traders tend to engage in discussions or conversations on investing information. Conversation, in the context of online investing, involves discussing alternatives, making predictions, asking questions, reporting observations, contributing opinions, sharing analysis and announcing decisions. Conversation is critical in the contagion of popular ideas about financial markets. People tend to pay more attention to ideas or facts that are reinforced

by conversations, rituals and symbols (Hirshleifer, 2001). De Long et al. (1990) stated that when opinions of noise traders are correlated with each other, they create risk. Furthermore, the noise trader literature further shows that the greater the number of noise traders there are in the market, the greater the increase in volatility (Koski et al. 2004). In short, noise traders have the ability to affect stock prices whenever information can be shared among investors and spread quickly through web channels (Zhang & Swanson 2010). Based on this understanding, I postulate that stock microblog sentiments extracted from online investor conversations have predictive power to influence future stock price movements. The conflict between efficient market and irrational investor views foreshadow the need for this study.

Research Questions

To discover the predictive value of stock microblogs over future stock price movements of both simple and market-adjusted returns, I put forward the following research questions:

1. What dimensions of stock microblogs (ticker-day, author-day or author-ticker-day) relate to high accuracy in predicting future directional stock price movements?
2. Which stock microblog feature sets are more salient towards predicting future directional stock price movements?

We answer these questions by proposing a design science approach that attempts to answer these questions by streamlining an approach to examine this predictive power at two levels: dimensions and feature sets, to understand the predictive relationship between stock microblog postings of Stocktwits.com, a nascent stock microblogging

channel, with future ten days of stock price movements. To answer the first question, I compare the predictive performances of three models based on three aggregated datasets, namely ticker-day, author-day, and the hybrid of author-ticker-day dimensions. To answer the second question, I examine five feature sets: namely author characteristics, investor sentiment, peer influence, peer network measures and market/ticker. The sample covers 360,000 microblog postings pertaining to 4570 different stock tickers from NASDAQ and NYSE stock exchanges, posted by 8935 distinct authors. I find relevant predictive power for both simple and market-adjusted returns, with high F-measures for the author-ticker-day dimension and sentiment and market/ticker feature sets. This study concludes that microblog sentiments do contain valuable information for investing decision making and supports the irrational investor hypothesis in influencing market prices.

I present the following three use cases to motivate the approach of using the three dimensions of author-day, ticker-day and author-ticker-day.

Let us examine the use case for the author-day dimension. An investor wishes to monitor the predictive accuracy of the top ranked individuals in the microblogging community. This ranking can be generated based on past performances of each individual based on number of tweets, past accuracy, number of followers, or any other measurable criterion. Figure 7 shows future 10 days predictive outcomes in relation to the individual's sentiment, bullish and bearish tweets for each individual in the list. The benefit of this information is that it allows others to observe the predictions of each individual in the community.

This approach aggregates microblog data by author for each market day. Its objective is to use features of authors to predict future outcomes. This approach takes advantage of author characteristics and author's microblogging measures as features in the classification models. Examples of these features are author demographics, trading preferences, suggested (expert label), total tweets sent, followers, following, and bullishness index. In addition, as in Chapter 3, I identified nascent measures based on author interactions which can account for peer influence, sentiment similarity and sentiment distance. However since investors are less likely to be diversified but instead focuses on a few stock indices (Huberman, 2001), aggregating all tweets for each author includes tweets with lower accuracy, hence the disadvantage of this approach. Authors tend to discuss tickers they are more familiar with and those tweets are more likely to have higher predictive power than tweets on unfamiliar stocks.

Let us next examine a use case for the ticker-day dimension (see Figure 8). An investor wishes to know the microblogging community's predictions of a few stock prices for the next 10 days. It shows the ticker, date of prediction, count of bullish and bearish tweets, sentiment (bullishness index) and outcome prediction for future 10 days. This list is filtered by total tweets and bullishness index (bullish or bearish). A correct prediction is the perceived prediction for investor to act upon while incorrect predictions should be ignored. For example, based on bullishness index of 1.83 and features of ticker and microblogs the system predict that stock price of M is going up 8 out of next 10 days. For those who are only concern of the short-term, then they should focus on predictions nearer to date of prediction (e.g., day 1 or 2).

This approach aggregates microblog data by ticker for each market day. Its objective is to predict outcomes for next 10 days. This approach takes advantage of ticker characteristics and ticker's microblogging measures as features in the classification models. Examples of these features are lagged closing prices, lagged trading volume, bullishness index, total tweets, average following, average followers and average message length. This dimension is also able to capture autocorrelation of past ticker performances with future outcomes that play a pertinent role in stock market data. However, since this approach ignores author characteristics, tweets from authors with low accuracy are likely to have an adverse effect on the predictive power.

And finally let us examine the use case for the author-ticker-day dimension (See Figure 9). An investor wishes to know the predictions for the author-ticker pair of author yCharts and M (Macy's) stock prices for the next 10 days. He search for author yCharts and select the link showing Figure 10. It shows the date of prediction, ticker, author, count of bullish and bearish tweets, sentiment (bullishness index) and outcome predictions for next 10 days. A correct prediction is the perceived prediction for investor to act upon while incorrect predictions should be ignored.

This approach aggregates microblog data by author and ticker for each market day. Its objective is to predict future 10 days outcomes for daily author-ticker pairs. This approach takes advantage of both author-day and ticker-day characteristics as features in the classification models.

Thus each author's tweets are grouped by ticker which should streamline the information to each ticker. This intuitively should result in higher accuracy per author-

ticker pair. The benefit of this dimension then is that it allows others to monitor predictions for specific author-ticker pairs.

These three use cases illustrate why the approach of examining dataset by dimensions can be very beneficial. The next step of model construction and evaluation of different feature sets further improves the prediction accuracy of our models. As stated by Shmueli and Koppius (2011, p. 3) on predictive analytics in IS research, "when the main purpose [of a model] is prediction but a certain level of interpretability is required, then predictive analytics can focus on predictors and methods that produce a relatively transparent model." A transparent model is one where constructs are easy to understand and built upon. I propose a set of predictive models to tease out the predictive power of different feature sets by adding each feature set to the model one at a time. This tease out technique is commonly used in data mining such as by Pant and Sheng (2013) where the scholars tease out different groups of web metrics in identifying competitor relationships. As compared to the more mainstream algorithmic feature reduction technique, the tease out technique is more transparent and intuitive. But it must be based on an understanding of the groups of features in the dataset. By testing each feature set separately and as an aggregation, I am able to determine the predictive power of each group of feature set by itself as well as in a combination with other feature sets.

Shmueli and Koppius (2011) state that research contributions of predictive analytics can be in the following: discovering new relationships, contributing to measures development, improving existing theoretical models, comparing existing theories, establishing the relevance of existing models and assessing predictability of empirical phenomena. This study has implications to the following three groups of stakeholders,

namely, researchers, investors and managers. On the research front, I established an approach to evaluate microblog features based on dimensions and features. This process of model construction and evaluation sets a benchmark for researchers and practitioners especially in other domains that are using microblogging channels extensively such as politics, marketing and health. As IS researchers, I study and explain the transformational impact of a nascent IT artifact (Agarwal & Lucas, 2005), stock microblogging, on the research and practitioner communities. Specifically I provide additional evidence of high predictability of UGC data to scholars of behavioral finance in supporting the tenets of irrational investor behavior in explaining stock market movements (De Long et al., 1990; Tetlock, 2007).

Related Work

Investor decision making outcomes are greatly impacted by actions and conversations from others around them (Adler & Adler, 1984; Blumer, 1975). “The emotions of fear and greed, coupled with subjective perceptions and evaluations of economic conditions and their own psychological predispositions and personalities, are major elements that affect the financial market behaviors” (Fung et al., 2005, p. 1). Based on this perspective I prioritize much of the initial literature review to focus on UGC and how it impacts economic outcomes for both stock related as well as retail channels. Subsequently, I review work in predicting stock market outcomes using textual analysis. Finally, I identify and discuss the research gaps that are addressed in this study.

UGC of Retail and Economic Outcomes

One active area of research from the marketing discipline relating consumer behavior to economic outcomes is closely associated with this study. It is popular for scholars in this area to study consumer behavior in the forms of customer reviews or electronic Word-of-mouth (eWOM) user ratings and blogs. The Internet's ability to reach a vast audience at low cost has placed new importance on Word-of-mouth (WOM) as a tool to influence and build trust (Dellarocas, 2003). I identify a few notable studies in this area.

Chevalier and Mayzlin (2006) examined effect of consumer reviews on sales of books for two online rivals, Amazon.com and BarnesandNoble.com. The authors confirm that customer WOM affects consumer purchasing behavior by observing the relationship between the number of reviews and average ranking with sales outcome. Duan, Gu and Whinston (2008) studied the relationship between online user reviews for movies and box office sales. They found that box office sales are influenced by the volume of postings, but the rating of reviews has no significant impact on sales. They concluded that businesses should focus more effort on facilitating consumer WOM exchanges instead of user ratings. Godes and Mayzlin (2004) tested Usenet WOM community interactions with ratings of new television shows. They found that WOM that is more dispersed across communities may be better than WOM that concentrates within each community.

UGC research in blogging, a closely related channel to microblogging, is a pertinent literature stream to examine. For example, Aggarwal et al. (2012a) investigate the influence of blog electronic WOM on venture financing. They discovered that eWOM of popular bloggers helps ventures in getting higher funding and valuations and that the

impact of negative eWOM is more than positive eWOM. In fact Aggarwal et al. (2012b) concluded that negative blogs may act as a catalyst that can exponentially increase readership. Similarly, Dhar and Chang (2009) investigated blog posts with sales of music and found that blog posts volume correlates positively with future sales of music CDs. Subsequently Fotak (2008) examined impact of blog recommendations on security prices and stock volume and Adamic and Glance (2005) examined political blogs and their impact on the 2005 election outcome.

On the microblogging front, Bollen, Pepe and Mao (2010a) used an extended version of Profile of Mood States (POMS) to extract six mood dimensions from over 9 million Twitter postings. They aggregated mood components on a daily scale and compared them to the timeline of cultural, social, economic and political events in the same time period. They found significant correlations between extracted mood dimensions and those occurring events. Bollen, Mao and Zeng (2010b) further expanded Bollen et al. (2011)'s study specifically towards predicting DJIA index over the same time period and conclude an accuracy of 87.6% and Mean Average Percentage Error (MAPE) by more than 6%. In another example, Asur and Huberman (2010) extracted sentiment from 6 million Twitter postings using LingPipe linguistic analysis package to predict box office revenue for movies. They benchmarked against Hollywood Stock Exchange (HSX) and obtained an accuracy of 0.94.

This sample of literature concludes that UGC does have significant positive correlation with economic outcomes. With this understanding of the extant literature comprehension, I next review UGC in virtual investing communities (VIC) to understand its relationship with regard to investment outcomes.

UGC of Virtual Communities

Virtual investing communities (VIC) are a popular source of social media for online investors. It seems to have blossomed in conjunction with the growth of the Internet. Its popularity stems from providing an environment where investors can collaborate and discuss, monitor what others are doing, or simply to seek fellowship (Wasko & Faraj, 2005). I review a few examples of studies undertaken to understand the relationship between behavior of community participants and stock market outcomes.

One of the earlier studies in this area is from Wysocki (1998), which used a sample of 3,000 stocks on a Yahoo! message board, and found that previous day returns, changes in trading volume, and changes in previous day postings have no predictive ability on stock returns. He did, however, find that an increase in volume of overnight postings led to a 0.18% average abnormal return. In addition, he concluded that total posting volume is higher for firms with high short-seller activity, extreme past stock returns and accounting performance, higher price earnings and book-to-market ratios, higher past volatility and trading volume, higher analyst following, and lower institutional holding (Wysocki, 1998). In another study, Tumarkin and Whitelaw (2001), using 181,000 postings from RagingBull.com found that, in general, message board activity does not predict industry-adjusted returns or abnormal trading volume. However they found that it is possible to predict the number of postings using previous day's trading volume, number of postings and weighted opinion (Tumarkin & Whitelaw, 2001).

A well-referenced paper, Antweiler and Frank (2004), using 1.5 million postings from Yahoo! Finance and RagingBull.com message boards, found significant but negative contemporaneous correlation between number of postings and stock returns on

the next day. The return, however, is economically very small in comparison to transaction costs. Nevertheless, message posting activity does help to predict volatility and trading volume. In addition, the authors concluded that volume of postings is positively correlated with volatility and bullishness. Similarly, Koski et al. (2004), apart from confirming that day traders are noise traders, also found that day-trading volume increases volatility but concluded no predictive relationship with stock returns. Das and Chen (2007) developed a methodology using five classifier algorithms to extract sentiment from stock message boards but found no significant predictive relationship between sentiment and stock prices. However, consistent with findings of Antweiler and Frank (2004), Das and Chen (2007) reaffirmed the existence of a significant correlation between posting volume and volatility but asserted that sentiment does not predict stock movements. Interestingly, Das et al. (2005) found that sentiment does not predict returns but instead returns drive sentiments. They implied that members of a virtual community are more likely to extrapolate past returns rather than to be contrarian, which ultimately leads to a behavior consistent with the representativeness heuristic (Das et al., 2005; Kahneman & Tversky, 1973; Lakonishok et al., 1994).

From the literature in the VIC stream, I conclude that UGC relationships with future outcomes are not conclusive. While some found correlation between features of UGC with financial indicators, both positive as well as negative, others were less successful.

Predictive Power of Textual Information

Predicting stock price movements using textual information via data mining and text mining is not new. Contrary to studies in the VICs, scholars in this area are able to find predictive power between media and stock market outcomes. However, the common data source used is news instead of UGC data, as used in this study. The following describes a few notable studies in this area.

One of the earlier work is from Wuthrich et al. (1998) predicting opening prices of five stock indices (DJIA, Nikkei, FTSE, Hang Seng, and Singapore Straits) by analyzing electronic news content of Wall Street Journal and other notable sources over a period of 3 months. Utilizing a k-NN learning algorithm, weights of relevant keywords and historical closing prices, the authors were able to obtain an average accuracy of 43.6%. Lavrenko et al. (2000) predicted intra-day stock price movements from real-time Yahoo business news stories using piecewise linear regression and language modeling techniques. Language modeling provides a framework for detecting stock price upward and downward trends. In a 40 day simulation, their average gain per transaction is .23%. Fung et al. (2005) predicted directional movements of stock price from content of real-time news stories from reuters using piecewise linear approximation algorithm. Their proposed system is able to achieve a hit-rate prediction accuracy of 65.4% at day 5. Hit-rate or stock price directional analysis is a measure of how often the sign of return is correctly predicted. This study utilizes the same directional analysis.

Robertson et al. (2007) predicted market reaction to stock specific news for S&P 100, FTSE 100, and ASX 100 indices. News is collected from 200 providers over a period of 5 months. They used terms of interest from Bloomberg news to predict

abnormal return and volatility using support vector machine (SVM) and C4.5 decision tree classifiers. They achieved over 80% accuracy, with best result from C4.5--88.25% accuracy. Schumaker & Chen (2009) is among the first to propose a system to predict numeric stock price movements in addition to directional movements. They used 9211 breaking financial news articles collected over a 5-week period covering S&P 500 stock indices. They used a combination approach of bag of words, name entities and noun phrases for text analysis and SVM for prediction classification. They obtained an overall accuracy of 57.1% and an overall return on simulation of 2.06%. This study utilizes a similar bag of words approach for text analysis.

Although much has been done in this research stream, there is a lack of effort in examining predictive power of microblogging.

Identifying Research Gaps

From the literature review I note that UGC has been a popular topic of research that is not exclusively in the retail domain but has permeated to the financial domain as well. However even though microblogging is fast gaining popularity, literature involving microblogging is limited as research using microblog is still in its infancy. Hence to the best of my knowledge, I note that there is a lack of research work correlating stock microblog features with stock market outcomes. Even more important is the fact that there is no clear process available as a baseline for future research and practice especially pertaining to best practice in examining predictive power of stock microblogging. I therefore seek to be among the first to embark on investigating features of stock

microblogs in understanding financial behavior phenomenon by establishing a process that examines both dimensions as well as features of stock microblogs.

Research and System Design

In order to address whether stock microblogs have predictive power over future stock price movements I propose a data mining classification approach to examine stock microblogs at two levels: dimensions and feature sets. See Figure 11.

Aggregating Stock Microblog Dimensions

In the text classification literature, classification can be conducted at the document, sentence or phrase level (Abbasi et al., 2008). Examples of document level classification are movie reviews (Pang et al., 2002), news articles (Wuthrich et al., 1998) and stock message web forums (Tumarkin & Whitelaw, 2001). Similarly sentence level classification is identification of subjectivity (Riloff et al., 2003) and phrase level classification is the identification of sentiment (Wilson et al., 2005). I adopt this approach to this study by examining the data at three levels which I term as the dimensions of ticker-day, author-day and author-ticker-day.

As in other classification studies, I examine the two-class problem, specifically whether the opinion of an author or average opinion for a ticker is accurate in relation to future stock price movements. All three dimensions are aggregated on a market day basis in following daily stock market activity. The intuition for the author-day dimension is that the feature sets of the author as well as microblog information for each author has predictive value. Similarly the intuition for the ticker-day dimension is that microblog

information from the masses pertaining to a single ticker may have predictive value. And finally the intuition for the author-ticker-day dimension combined features of both author and ticker/market that is postulated to have salient predictive power.

Extracting Stock Microblog Feature Sets

I incorporate past literature to decide on the stock micro feature sets to group. Although each dimension may not have all feature sets, the description for each feature set is the same if it exists in that dimension. The feature sets are 1) author sentiment/opinion, 2) author characteristics, 3) peer influence, 4) peer network measures and 5) market/ticker measures.

Investor Sentiment/Opinion

Stock microblog posting message limit constraint of 140 characters is advantageous as it forces users to write succinctly using meaningful keywords. This leads to low time investment and high frequency of generated postings (Java et al., 2009). It further increases the density of useful information and those keywords are more likely to be repeated by others. These keywords may contain opinions or sentiment from the author pertaining to the prediction of future stock performance. For example postings see Table 13.

Sentiment is a key predictor in this study. Since sentiment is derived from text and are not provided by the authors, I have to either manually or automatically extract them. Although human labeling is the better alternative, due to resource constraint I

resort to systematically labeling the large volume of postings. This process is explained in Chapter 2.

Author Characteristics

Prior studies in various fields have shown that source information has direct impact on future outcomes such as sales (Chevalier & Mayzlin, 2006), information quality (Oh & Sheng, 2013) and peer influence (Aral, 2011). Thus I include them as the author feature set in this study. Author features are extracted from two sources: author profiles available on the microblogging channel Stocktwits.com, as well as from the microblog content of each author. I aggregate demographic information such as name, location and bio and trading preferences such as investment approach, risk and holding strategy into binary variables of demographic and trading self-disclosed variables implying the presence or absence of such information for each type. Other variables for the author are average posting time of the day (TOD), average posting day of the week (DOW), market (NYSE or NASDAQ), number of followers and followings, suggested (expert label), average message length of each posting and total postings sent.

Peer Influence

Financial market uncertainties and risk of adoptions drive investors to seek the opinion of others in their decision making process (Becker, 1970; Cancian, 1979) introducing the notion of peer influence (Aral et al., 2009). Peer influence, also known as social influence (Putnam, 1993; Rice et al., 1990), social contagion (Iyengar et al., 2011; Susarla et al., 2012) or peer effects (Bandiera & Rasul, 2006; Sacerdote, 2001), refers to a

phenomenon whereby an actor's decision on the adoption of a new product (or behavior) is dependent on other actors' attitudes, knowledge, or adoption (Susarla et al., 2012; Van den Bulte & Lilien, 2001).

In measuring peer influence, I first adhere to Goldenberg et al. (2009) and Trusov et al. (2008) by defining links by activity (e.g., retweets, mentions) and not by pointers (e.g., follower) as a pointer between two individual in a social networking site does not necessary imply influence (Goldenberg et al, 2009). Second, I adopt Iyengar et al. (2010) in using indegree nominations of peers as a measure of peer influence. According to Iyengar et al. (2010), people who are often nominated by peers as someone they turn to for expertise or discussion are likely to be true sources of influence. Based on these two assumptions, I selected two groups of peer influence measures as the peer influence feature set: baseline indegree counts (RT_in, mention_in and reply_in) and residual measures (indegree-outdegree).

For the first group, I adopt two basic indegree and outdegree measures from the literature, namely number of retweets and mentions. These measure the incoming and outgoing actions among members of the community. The measures are retweet in (RT_in) and out (RT_out), and mention in (mention_in) and out (mention_out).

For the second group, I account for the effect of reciprocity as explained by Weng et al. (2010) by subtracting outdegree from the indegree measures (e.g., indegree – outdegree) resulting in residual measures (e.g., RT_diff = RT_in – RT_out). Agarwal et al. (2008) defined this as InfluenceFlow. The intuition is that the more inlinks a blog post acquires the more recognized it is while an excessive number of outlinks jeopardizes the novelty of a blog post. Thus an individual with many indegree accompanied by many

more outdegree should have a lower influence compared to another with high indegree but lower outdegree. Peers are obligated to reciprocate thus sheer number of outdegree naturally begets high number of indegree. Despite its novelty and intuitiveness, it is surprising that few research studies in microblogging adopt this measure.

Peer Network Measures

For peer network feature set, I measure how relevant is an individual to his/her peers by accounting for the proportion of attention given by the individual's peers, both in terms of peers' outdegree count as well as peers' number of peers. The intuition is that an individual with more peer attention should be more influential. I based this on the TF-IDF (term frequency – inverse document frequency) concept (Salton & McGill, 1989) from information retrieval. Term frequency (TF) (Wu et al., 2008) refers to how relevant a word (term) is in a collection or corpus. These features are operationalized below.

Normalized influence by peer outdegree (NIPO) = (indegree to A) / (outdegree from A's peers). An example is provided in Appendix (See page 170).

Market/Ticker Indicators

Since autocorrelation of past and future stock prices is a very important element in the stock market (Wu et al., 2008), it makes good sense to use past ticker and market such as past 5-days (lagged) stock ticker closing prices, Dow Jones Industrial Average and stock ticker volume as features in this study. The same consensus in the field of finance verifies this close correlation between past market indicators with future financial

performance (Antweiler & Frank, 2004; Sabherwal et al., 2008; Wysocki, 1998). The selected measures are 5-day lagged measures as described below.

Dow Jones Industrial Average Index is a measure of the performances during a standard trading session of 30 large publicly owned companies in the US (Wikipedia, 2013). It is one of the most closely watched benchmarks monitoring stock market activity.

Stock ticker volume is the number of shares traded during a particular trading session. Essentially it represents the amount of shareholder interest in the stock index and the level of liquidity of the company.

Stock ticker closing price is the last traded price of the index during a particular trading session. A stock's closing price can be compared to the previous day's closing price or the day's opening price. A stock is said to have closed up, or higher, when it closes above the previous day's close, and down, or lower, when it closes below the previous day's close. A stock that closes higher than it opened shows strength; a stock that closes lower than it opened shows weakness.

System Design

I propose the following system design (shown in diagram below) with its accompanying four major steps: pre-processing, dimension creation, feature extraction, and classification (dimension classification followed by feature classification). See Figure 12.

The downloaded microblog postings are first pre-processed. I discard postings without any ticker, more than one ticker, or those not in NASDAQ and NYSE exchanges

(see Appendix page 186). Then I aggregate postings and author information into ticker-day level, author-day level and author-ticker-day dimension. Any observations with number of posts per period fewer than 3 are discarded to reduce noise. Subsequently, for each dimension I extract all relevant features as per the feature sets discussed in previous sections. Finally, I perform classifications on first the dimension data and then the feature set data. Evaluation and analysis followed thereafter.

Data and Evaluation

Data

The data in this study are from two sources: stock microblog posts is from Stocktwits.com, a popular stock microblogging channel and stock inter-day trading data is from Google Finance, for the period from Nov 1, 2011 to April 18, 2012. After initial pre-processing removing nonrelevant posts I obtained over 360,000 postings from 8935 authors pertaining to 4570 stock tickers. The top 20 stock tickers by stock microblog volume are responsible for over 40% of all the posts. See Table 14. This is consistent with prior finding that people often invest in familiar stocks and they tend to ignore principles of portfolio theory (Huberman, 2001).

Feature Variables

Table 15 lists all feature variables grouped by each respective feature set. Some of these variables are known while others are extracted. This grouping of feature set was determined by performing a feature selection using Weka data mining feature selector InfoGainAttributeEval evaluator and Ranker search method. The ranked list of features

(see Appendix page 188) showed features grouped by market/ticker, sentiment, author, peer influence and network measures in that order.

Class Variables

The class variable in this study is the predictive outcome matching author average bullishness index of each observation with future 10 days stock price movements. There are two types of class variable, one that is for simple return which refers to the raw return between close price and open price of a stock ticker. Another is the market return which refers to whether the simple return is better than the market return. While the simple return is popular with stock market prediction studies as an outcome variable, the market-adjusted return is more robust and accounts for the stock's ability to perform better than the market. The details are discussed below:

Simple Return Class Variable

This is a binary (1,0) prediction outcome measure matching bullishness index (opinion/sentiment) from ticker level aggregated microblogs, where bullish (>0) or bearish (<0), against actual stock price movement, upward (1) or downward (-1) trend. The value 1 is for correct prediction, 0 for incorrect prediction. Each observation has 10 class variables, representing future 10 days, being period *(t+1)* to *(t+10)*.

Market-Adjusted Return Class Variable

This is a binary (1,0) prediction outcome measure as described in the simpleSimple return measure above but accounts for market return (Dow Jones Industrial Average). Specifically, the system matches ticker level sentiment against the net trend

between simple return and market return. The value 1 is for correct prediction, 0 for incorrect prediction. Each observation has 10 class variables, representing future 10 days, being period *(t+1)* to *(t+10)* (see Appendix page 187 for an example).

Performance Measures

In line with standard metrics from information retrieval (Ma et al., 2009), I report precision, recall and F-measure scores in evaluating the performance of all predictive models. Description of each measure is given below.

Precision is a measure of exactness; specifically it is the fraction of predicted instances that are relevant (Equation 8).

$$Precision = \frac{Number\ of\ correctly\ predicted\ positive\ (negative)\ instances}{Number\ of\ predicted\ positive\ (negative)\ instances} \quad \text{(Equation 8)}$$

Recall is a measure of completeness; specifically it is the fraction of relevant instances that are successfully retrieved (Equation 9).

$$Recall = \frac{Number\ of\ correctly\ predicted\ positive\ (negative)\ instances}{Number\ of\ actual\ positive\ (negative)\ instances} \quad \text{(Equation 9)}$$

F-measure is a harmonic mean measure that combines precision and recall where both are equally weighted (Equation 10).

$$F\text{-}measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad \text{(Equation 10)}$$

Dimension Models

In answering the research question which dimension has the highest prediction accuracy, I aggregate stock microblogs into three models as per each dimension: author-

day (D1), ticker-day (D2) and author-ticker-day (D3). Due to the differences between dimensions, each model might have different set of features as compared to another. Priors for the three models are in Appendix (see Appendix pages 194-196). Comparatively on average there are significantly more bullish than bearish observations (e.g., D3 .804 bullish, .196 bearish). And bullish also has higher accuracy than bearish (e.g., D3 .533 bullish, .446 bearish). Interestingly, while accuracy for bullish is increasing over time, accuracy for bearish is reducing, from t1 to t10.

Feature Set Models

In answering the research question which feature set(s) has the highest prediction accuracy, I examine 11 models: M1 to M11, each having one or more combination of feature sets. By doing so predictive power of different features sets are distinctively observed. The main Model M1 is the de facto model encompassing all features. Models M2 to M6 contain a single feature set each. And models M7 to M11 contain a combination of feature sets within each model (see Table 16).

Experiment 1: Examining Microblog Dimensions

I conduct Experiment 1 to examine which microblog dimensions relate to high predictive accuracy over future stock price directional movements. Classification tests are conducted on three models (D1, D2 and D3) based on the three dimensions of ticker-day (D1), author-day (D2) and author-ticker-day (D3) using 10-fold cross validation via eight classifiers: Naïve Bayes, Logistic, ZeroR, Random Forest, SMO, AdaBoost, Bagging (J48) and ClassificationViaRegression (CVR) from the Weka data mining package

(Witten & Frank, 2005). These classifiers are selected from the main classifier groups namely tree classifier, Bayesian and regression because they are popularly used in testing a wide array of applications and domains (Witten & Frank, 2005). Results in Table 17 show that D3 (author-ticker day dimension) has the highest average F-measure accuracy (e.g., Bagging F=.884 simple, F=.882 market and Random Forest: F=.879 simple, F=.876 market). On the other hand, both D1 and D2 performed much worse than D3 with ticker-day (e.g., Bagging F=.638 simple, F=.629 market and RF F=.616 simple, F=.608 market) better than author-day (e.g., Bagging F=.455 simple, F=.464 market and RF F=.557 simple, F=.558 market). Results for precision and recall are similar and they are reported in Tables 18 and 19, and Appendix (see Appendix pages 190-192). Performances for Bagging, Random Forest and classification via regression classifiers are very similar. Tree and regression classifiers seem to be suitable for microblog features as they are more robust to the presence of noise and run efficiently on large datasets (Caruana et al., 2008). Zero-R has the worst result showing that microblog features do contain predictive power that can be explained beyond the simple majority rule approach.

Experiment 2: Examining Microblog Features

From dimension classification I then proceed to feature classification. Based on the results of Experiment 1, I use the author-ticker-day model (D3) to examine which feature sets relate to high predictive accuracy over future stock price directional movements. Classification tests are conducted on feature set models: M1 to M11, using 10-fold cross validation. With both Random Forest and Bagging (J48) being the better classifiers, I selected Random Forest since it is computationally more efficient than

Bagging. Each model has 10 days future predictive outcomes (t1 to t10) and pertains to two class variable types: simple and market return. The defacto M1 model with all feature sets generated a mean F-measure of .889 simple (F=.888 market). However, when I individually classify each group of feature set (M2 to M6), I discovered that the best single feature set predictor is market/ticker information (F=.77 simple, F=.768 market) followed by author (F=.717 simple, F=.715 market) and sentiment (F=.518 simple, F=517 market). Then I proceed with the aggregated models and found that M8 model with feature sets of market/ticker, author and sentiment having the best F accuracy (F=.901 simple, F=.902 market). Hence the reason why M1 is worse off than M8 is due to the noisy features of peer influence and network. Results are shown in the following Table 20. Results for precision, recall, accuracy, class-0 F-measure and class-1 F-measure are reported in Appendix (See Appendix pages 200-202), the model with all but market-ticker features has impressive accuracy (F=.729 simple, F=.729 market) showing that microblog features alone already have high predictive power. Logically due to the missing critical features in each dimension, predictive power of both sets is lower than that of D3. As an exploratory extension, I regroup features of network feature set to sentiment and peer influence feature sets and this resulted in significant performance improvements for both those feature sets. Results for both extended M3 (F=.551 simple) and extended M5 (F=.477 simple) has improved. See Table 21. Results for other performance measures for testing M1 to M11 are provided in Appendix (see pages 197-202).

Results Discussions

First, I find that the author-ticker-day (D3) dimension has the highest predictive accuracy over future stock price movements. This is because it incorporates both author and market/ticker dimensions. Hence, the opinions (sentiment) of more-accurate authors and less-accurate authors per ticker are separated allowing the model to perform better.

Clearly as shown in the D1 and D2 dimensions where either ticker or author information are respectively missing, a significant portion of the predictive power is also removed when these feature sets are absent. Seeking the right dimension model to aggregate microblog information is a basic but critical decision in this classification process.

Subsequently for feature set, I find that market-ticker indicators have high predictive power, a finding that is consistent with the extant literature (Antweiler & Frank, 2004). However, the more interesting findings are that author characteristics and opinions (sentiment) of microblogging authors also have predictive power and the aggregate of all three feature sets have the highest predictive power. Market-ticker features have high predictive power due to auto-correlation effect from market dynamics. This is a pertinent factor to consider when evaluating data relating to the stock market. Simple and market returns are similar in most comparisons and the proposed models are robust to cater for both types of outcome variables.

The saliency of author features show that source characteristics (Forman et al., 2008) play an important role in the predictive value of microblogs. This supports the conclusion of Chapter 3 of this dissertation where certain author characteristics such as self-disclosed trading preferences and number of total tweets are significant in explaining

information quality. Individuals with high accuracy have the experience and expertise which are captured in the author feature set. In addition, their microblogging behavior relating to different tickers also contribute to the predictive power. Hence the need to aggregate this data by the dimension of author and ticker. However this alludes to the opportunity for performing feature reduction for these models to further improve and discover which features are more salient in each model.

The effect of investor sentiment is marginal to the predictive power. This proves that opinion alone is not enough to aid prediction since opinions from individuals are not reliable and changes rapidly. Using simple mean comparison from priors for this dataset I note that although majority of the sentiment is bullish, bearish posts are more accurate. Behavioral influences such as wishful thinking (Seybert & Bloomfield, 2009), overconfidence (Barberis & Odean, 2001) and the effect of negative information (Aggarwal et al., 2012b; Luo, 2007) may play a part in this phenomenon.

However author feature sets of peer influence and network measures do not seem to have any predictive power. This is probably due to low frequency of these measures. In Chapter 3, I also found that peer influence has a negative correlation with accuracy and has a low R-squared.

Microblog UGC does have features with predictive power. However, the need to have the correct framework emphasizing different dimensions and features is salient in relating to this predictive power. Hence the framework proposed in this study accomplished that objective.

Conclusions and Future Directions

A recent Wall Street Journal article highlighted a pertinent change in the investment climate. “More and more investors are not poring through corporate reports searching for gems and duds, but are trading big buckets of stocks, bonds and commodities based on mainly macro concerns (big picture market movers like the economy, politics and regulation)” (Lauricella & Zuckerman, 2010). Since news of such macro forces diffuse profusely through social media such as stock microblogging channels (Antweiler & Frank, 2004), this study of investigating impact of investor sentiment on stock performance is imperative. It makes two important contributions: first to the research community and second to practitioners.

In conclusion, microblog content, particularly investor opinion/sentiment, does appear to have strong predictive value for future market directions. I conclude this by studying sentiment from 360,000 microblog postings from Stocktwits.com, a stock microblogging service, over a period of 3 months. The principal contribution of this study is to present an approach to analyze microblog features for future research utilizing microblogging domains such as those from marketing, politics, health and social studies. In so doing I heed the call of Agarwal and Lucas (2005) in explaining the transformational impact of a nascent IT artifact, stock microblogging, in connecting to reference disciplines. Specifically, I provide evidence for the model of irrational investor sentiment, recommend a supplementary investing approach using user-generated content (UGC) for investors and a framework that may contribute to the monetization schemes for Virtual Investing Communities (VIC) for managers.

Contribution to Research Community

I propose a model construction and evaluation to extract relevant dimensions and features from stock microblogs in relating to future outcomes. This approach is applicable to other domains that are present in the microblogging channel such as marketing (advertising, promotions, etc.), politics (campaigns, election debates, etc.) and health (outbreaks). It serves as a benchmark for future work to explore predictive power of microblog.

Agarwal and Lucas, Jr. (2005) exhort IS scholars to initiate research with a greater macro focus in establishing a stronger IS credibility and identity. I heed this call in studying and explaining the transformational impact of a nascent IT artifact, stock microblogging in strengthening ties with reference disciplines. This study is especially applicable to the debate between efficient market and behavioral finance in supporting the latter in its tenet that investor sentiment does have influence on stock market prices (De Long et al., 1990; Tetlock, 2007). In addition, I established the presence of the predictive power of microbloging. Weblog. I also add to the research stream investigating relationship between artifacts in online communities, specifically UGC data, and economic outcomes such as stock market (Antweiler & Frank, 2004) or retail sales (Godes & Mayzlin, 2009). This opens up many future research opportunities for scholars who are seeking to understand the impact of UGC on sales and other economic outcomes.

Contribution to Practitioners

There are two groups of practitioners addressed by this study: investors and managers. Investors both personal and institutions, have always been searching for

effective techniques to predict stock outcome. This study adds another dimension to the search by uncovering the predictive value of UGC sentiment. This approach provides the investor community with a more scientific approach to make informed and precalculated decisions in stock investing. This is in line with the study's contribution to platform providers such as Stocktwits, whereby it sheds more light for managers to enhance their filtering and search mechanism. This allows their users to overcome limitations posed by information overload and inadequate cognitive capacity by focusing on a smaller group of relevant experts or postings to monitor. The three use cases in the beginning of the Chapter are a case in point. Sentiment may also be a viable criterion for hedge-fund managers to evaluate candidate indices to be included in their portfolios.

Furthermore, managers may opt to design investing tools to enable investors in analyzing stock postings. For monetization purposes they may also incorporate advertisements with postings deemed with higher predictive value or even establish partnerships with selected authors of predictive postings. These findings may be valuable for UGC channels such as Twitter which has yet to establish any significant revenue stream (Miller, 2010).

Future Directions

An important feature to examine is the magnitude of stock price change since in this study I only test directional accuracy of stock movements. This could expand to include categories of stock price change from very large decrease in price to very large increase. This question is important as even with a model that gets 80% of predictions correct, 20% wrong predictions could lead to adverse outcomes. This scenario could

vastly undermine the value of the underlying model. In addition, it would be interesting to explore narrower time ranges for predictions. Time ranges of minutes to hours after the posting could yield fascinating results. Long-term time ranges such as weeks may also be important to examine as well.

One important extension of this study is to expand the features to cover other source of public and private information to include news, analyst forecasts, stock message boards and financial blogs. This may or may not strengthen the predictive power of stock microblog sentiment. It is interesting to understand how these external sources impact the activities and predictive accuracy of the community. I could also include more stocks of different characteristics (small and large capitalization, etc.) as well as other investing instruments such as foreign exchange and futures to further enrich the dataset and discover deeper relationships between predictive accuracy stock microblogging and characteristics of investment performance.

Another extension is to propose an adaptive learning algorithm to enhance the performance of classifiers by removing instances that are of lower predictive quality. I plan to adopt the approach of Abbasi et al. (2010) where the scholars used adaptive learning in their algorithm “Recursive Trust Labeling” in classifying/identifying fake medical websites. The intuition is to apply knowledge gained from classification of stock microblog dimensions in past period dataset to the current or future period datasets. I note that the combination of author-ticker pair is usually a good dimension to observe investing information. In short an author who tweeted about certain ticker and was correct on average in the past is likely to also be correct on the same ticker in the future. There is an inherent investing knowledge tied between author-ticker pairs.

Suppose investor A tweets about GOOG (Google) and BAC (Bank of America) while B tweets about BAC and APPL (Apple). See Figure 13. Suppose the prediction accuracy is as such:

(A,GOOG) = 1

(A,BAC) = 0

(B,BAC) = 0

(B,APPL) = 1

In the next period, the likelihood of (A,GOOG)=1 is higher than (A,BAC)=1. Similarly (B,APPL) > (B,BAC). This is true as investors devoted much time and focus on certain stock tickers that he/she developed expertise in and intuition regarding their stock trends. So people who were accurate in the past are likely to be accurate again in the future. Thus, eliminating low accuracy pairs such as (A,BAC) and (B,BAC) are (intuitively) likely to increase classification accuracy of future models.

Leveraging the knowledge gained from Abbasi et al. (2010) I outline this adaptive learning algorithm in the Figure 14. The basic intuition is that predictive value of instances (author-ticker pairs) from the initial classification can be ranked. And the lower quality group to be applied and removed from the next dataset.

I train the first model (as in D1-D3) then I test the second model. Based on the classification results of the first model, I remove false negative and false positive of author-ticker pairs found in the second model. Intuitively the same author-ticker pairs that are accurate in model 1 should be accurate in model 2, on average. This approach should increase the accuracy of these models and also provide usability options for users/investors in identifying favorable author-ticker pairs to follow (see Figure 14).

Table 13. Examples of microblog posts with bullish/bearish sentiment

Keywords	**Example Posting**	**Sentiment**
breakout	$CLW nice **breakout** this morning	Bullish
shorting	**Shorting** $Amzn 300 pieces @ 131	Bearish
pop	$RIMM RIM announcing new phone today 11ET, there could be a **pop**	Bullish
Keep eye	http://chart.ly/qy3ph3 $AAPL - **keep eye** on this one today	Bullish
bearish	Intel Turns **Bearish**: http://drduru.com/onetwentytwo/2010/07/31/intel-turns-bearish/ $INTC	Bearish
cuts	$TXN: Robert W. Baird **cuts** to Neutral	Bearish
loss	$MNKD reported just a little bit ago - **loss** of 0.37 per share vs. estimates of a 0.40 loss per share... Watch it for an intraday move.	Bearish
Accumulate, top	$C continue to **accumulate** under 4.5 may be a **top** pick of the year. analyst proj 5.75 by YE	Bullish

Table 14. List of top tickers by microblog volume

Ticker	Company Name	Microblog Count	Percentage to all dataset
AAPL	Apple	72629	19.82655
NFLX	Netflix	7151	1.952108
GOOG	Google	6231	1.700963
AMZN	Amazon	5753	1.570476
RIMM	Blackberry	5525	1.508236
SINA	Sina Corp	5488	1.498136
BAC	Bank of America	4844	1.322334
GMCR	Green Mountain Coffee	4107	1.121145
BIDU	Baidu	3800	1.037339
FIO	Fusion-io	3791	1.034882
RENN	Renren	3287	0.897298
VVUS	Vivus	3278	0.894841
PCLN	Priceline	2871	0.783737
GLD	SPDR Gold	2846	0.776912
GS	Goldman Sachs	2779	0.758622
ZNGA	Zynga	2773	0.756984
DMND	Diamond Foods	2610	0.712488
LNKD	LinkedIn	2466	0.673178
LGF	Lions Gate Entertainment	2360	0.644242
SODA	Soda Stream	2330	0.636052

Table 15. Feature set descriptions

Feature Set	Feature	Known/Extracted	Values	Description
Author	TOD	Extracted	1-24	Time of the Day (1-12 am till 1 am, 2-1am till 2 am, etc).
	DOW	Extracted	1-7	Day of the week (1-Monday, 2-Tuesday, etc).
	market	Extracted	0 or 1	Is the posting submitted during trading hours?
	follower	Known	Numeric	Those who follow this author.
	following	Known	Numeric	Those who this author follows.
	demographic_disclose	Extracted	0 or 1	Any demographic variable exists?
	trading_disclose	Extracted	0 or 1	Any trading preference variable exists?
	suggested	Known	0 or 1	Is the post from an expert?
	avg_msg_len	Extracted	Decimal	Average Count of words in the posting.
	total_tweets	Known	Numeric	All postings ever sent by author.
Peer Influence (Microblog)	RT_in	Extracted	Numeric	Indegree counts of retweets by other individuals of this individual's tweets
	RT_out	Extracted	Numeric	Outdegree counts of retweets by this individual of other individuals' tweets
	RT_diff	Extracted	Decimal	Residual of indegree and outdegree retweets
	mention_in	Extracted	Numeric	Indegree counts of replies from other individuals to this individual
	mention_out	Extracted	Numeric	Outdegree counts of replies from this individual to other individuals
	mention_diff	Extracted	Decimal	Residual of indegree and outdegree mentions
Sentiment (Microblog)	bullish_index	Extracted	Decimal	Measure of bullishness aggregated over all tweets posted by the investor
	disagree_index	Extracted	Decimal	Measure of polarity of the tweets sentiment
Market	dow1	Known	Numeric	Dow Jones Industrial Average Index – representing market performance.
	dow2	Known	Numeric	Dow Jones Industrial Average Index
	dow3	Known	Numeric	Dow Jones Industrial Average Index
	dow4	Known	Numeric	Dow Jones Industrial Average Index
	dow5	Known	Numeric	Dow Jones Industrial Average Index

Table 15 Continued.

Feature Set	Feature	Known/Extracted	Values	Description
Ticker	ticker_close1	Known	Decimal	Ticker closing price for *t-1*.
	ticker_close2	Known	Decimal	Ticker closing price for *t-2*.
	ticker_close3	Known	Decimal	Ticker closing price for *t-3*.
	ticker_close4	Known	Decimal	Ticker closing price for *t-4*.
	ticker_close5	Known	Decimal	Ticker closing price for *t-5*.
	ticker_vol1	Known	Numeric	Ticker volume for *t-1*.
	ticker_vol2	Known	Numeric	Ticker volume for *t-2*.
	ticker_vol3	Known	Numeric	Ticker volume for *t-3*.
	ticker_vol4	Known	Numeric	Ticker volume for *t-4*.
	ticker_vol5	Known	Numeric	Ticker volume for *t-5*.
Peer Network Measures (Microblog)	RT_norm_influence_peer_ outdegree	Extracted	Decimal	Normalized of indegree retweets by peers' outdegree
	RT_sentiment_similarity	Extracted	Decimal	Similarity measure of author and his/her peers, in the RT network
	RT_sentiment_distance	Extracted	Decimal	Average sentiment measure between author and his/her peers, in the RT network
	mention_norm_influence_peer_ outdegree	Extracted	Decimal	Normalized of indegree mentions by peers' outdegree
	mention_sentiment_similarity	Extracted	Decimal	Similarity measure of author and his/her peers, in the mention network
	mention_sentiment_distance	Extracted	Decimal	Average sentiment measure between author and his/her peers, in the mention network

Table 16. Models with different feature sets

Model	Feature sets included in the model
M1	All feature sets are included.
M2	Only market/ticker feature set.
M3	Only sentiment feature set.
M4	Only author feature set.
M5	Only peer influence feature set.
M6	Only network feature set.
M7	Only market/ticker AND author feature sets.
M8	Only market/ticker, author AND sentiment feature sets.
M9	Only market/ticker, author AND network feature sets.
M10	Only market/ticker, author AND peer influence feature sets.
M11	All but market/ticker feature set.

Table 17. Average F-measure result comparing dimensions D1, D2 and D3

DV	Classifier							
Simple	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
D1	0.440	0.507	0.359	0.616	0.443	0.572	0.638	0.621
D2	0.551	0.500	0.455	0.557	0.564	0.456	0.455	0.501
D3	0.541	0.562	0.362	0.879	0.539	0.573	0.884	0.840
Market								
D1	0.442	0.504	0.361	0.608	0.440	0.575	0.629	0.613
D2	0.555	0.502	0.464	0.558	0.567	0.465	0.464	0.516
D3	0.539	0.562	0.358	0.876	0.545	0.571	0.882	0.838

Note: 1-NB, 2-Logistic, 3-ZeroR, 4-RF, 5-SMO, 6-AdaBoost, 7-Bagging, 8-CVR

Table 18. Average precision result comparing dimensions D1, D2 and D3

DV	Classifier							
Simple	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
D1	0.536	0.533	0.273	0.622	0.522	0.589	0.638	0.621
D2	0.553	0.556	0.365	0.558	0.561	0.421	0.365	0.561
D3	0.554	0.566	0.276	0.882	0.559	0.608	0.885	0.840
Market								
D1	0.533	0.532	0.275	0.614	0.519	0.589	0.629	0.614
D2	0.556	0.560	0.374	0.558	0.564	0.424	0.374	0.572
D3	0.554	0.565	0.272	0.879	0.562	0.604	0.882	0.838

Note: 1-NB, 2-Logistic, 3-ZeroR, 4-RF, 5-SMO, 6-AdaBoost, 7-Bagging, 8-CVR

Table 19. Average recall result comparing dimensions D1, D2 and D3

DV	**Classifier**							
Simple	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
D1	0.530	0.538	0.522	0.621	0.534	0.589	0.638	0.622
D2	0.580	0.602	0.604	0.583	0.578	0.604	0.604	0.604
D3	0.546	0.567	0.525	0.879	0.556	0.596	0.884	0.840
Market								
D1	0.531	0.537	0.524	0.614	0.535	0.590	0.629	0.615
D2	0.585	0.609	0.611	0.587	0.585	0.611	0.611	0.612
D3	0.546	0.566	0.522	0.876	0.558	0.591	0.882	0.838

Note: 1-NB, 2-Logistic, 3-ZeroR, 4-RF, 5-SMO, 6-AdaBoost, 7-Bagging, 8-CVR

Table 20. Classification results for models M1 to M11.

Simple	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
F-measure	0.889	0.770	0.518	0.717	0.418	0.486	0.826	0.901	0.826	0.825	0.729
Accuracy	0.889	0.770	0.547	0.717	0.529	0.549	0.826	0.901	0.827	0.826	0.730
Precision	0.891	0.770	0.547	0.718	0.529	0.561	0.828	0.902	0.829	0.828	0.731
Recall	0.889	0.770	0.547	0.717	0.529	0.549	0.826	0.901	0.827	0.826	0.729
class-1 F	0.888	0.777	0.599	0.716	0.574	0.600	0.825	0.901	0.825	0.824	0.725
class-0 F	0.889	0.756	0.422	0.716	0.234	0.350	0.826	0.900	0.827	0.826	0.733
Market	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
F-measure	0.888	0.768	0.517	0.715	0.413	0.486	0.825	0.902	0.826	0.825	0.729
Accuracy	0.888	0.769	0.541	0.716	0.526	0.548	0.825	0.902	0.826	0.826	0.729
Precision	0.890	0.769	0.541	0.717	0.526	0.561	0.827	0.903	0.828	0.828	0.732
Recall	0.888	0.769	0.541	0.716	0.526	0.548	0.825	0.902	0.826	0.826	0.730
class-1 F	0.887	0.781	0.578	0.712	0.525	0.567	0.823	0.901	0.823	0.823	0.727
class-0 F	0.889	0.757	0.442	0.717	0.276	0.387	0.826	0.902	0.828	0.827	0.730

Table 21. Simple return F-measure for extended M3 and M5

Day	Extended M3	Extended M5
1	0.537	0.493
2	0.545	0.492
3	0.544	0.519
4	0.534	0.504
5	0.553	0.460
6	0.554	0.462
7	0.553	0.465
8	0.562	0.456
9	0.563	0.455
10	0.565	0.463
Mean	0.551	0.477

Based on date: 3/4/2013

Stock price directional Predictions for 10 days. Based on author-day Information.

Author	Bullish	Bearish	Sentiment	Prediction is likely to be									
				1	2	3	4	5	6	7	8	9	10
chartly	24	3	1.833	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗
John	20	1	2.351	✗	✓	✓	✓	✗	✓	✓	✓	✓	✓
1nvest	17	5	1.099	✓	✗	✓	✓	✗	✓	✗	✓	✗	✓
Mama	19	6	1.050	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
daytrad	20	5	1.253	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓
zeroguy	16	8	0.636	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
HCPG	14	9	0.405	✓	✓	✗	✓	✓	✓	✗	✓	✓	✗
Ycharts	12	12	0.000	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
Ronin	10	15	-0.375	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓
djtrade	9	16	-0.531	✗	✓	✓	✓	✓	✓	✓	✗	✗	✓

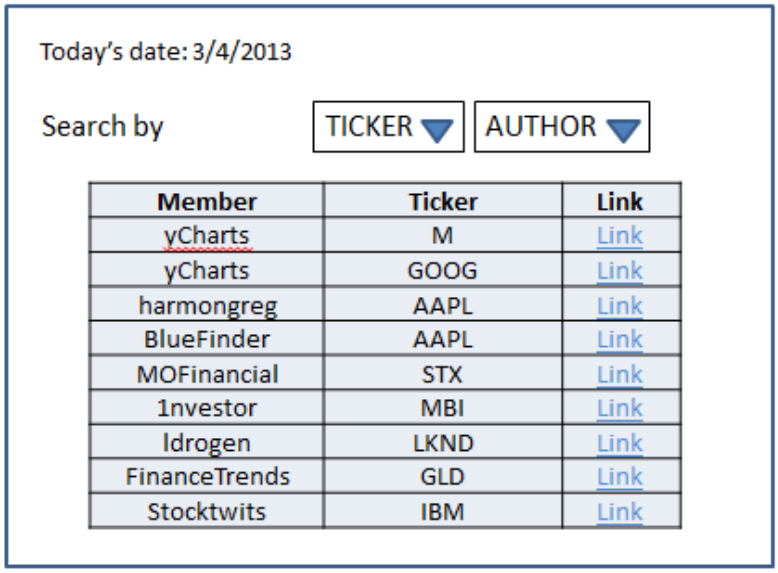
Figure 7. Stock price directional predictions for author-day dimension

Based on date: 3/4/2013

Stock price directional Predictions for 10 days. Based on Community and market Information

Ticker	Bullish	Bearish	Sentiment	Prediction is likely to be									
				1	2	3	4	5	6	7	8	9	10
M	24	3	1.833	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗
GOOG	20	1	2.351	✗	✓	✓	✓	✗	✓	✓	✓	✓	✓
C	17	5	1.099	✓	✗	✓	✓	✗	✓	✗	✓	✗	✓
BAC	19	6	1.050	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
STI	20	5	1.253	✗	✓	✓	✓	✓	✗	✓	✓	✓	✓
MBI	16	8	0.636	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓
LKND	14	9	0.405	✓	✓	✗	✓	✓	✓	✗	✓	✓	✗
Y	12	12	0.000	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓
AAPL	10	15	-0.375	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓
A	9	16	-0.531	✗	✓	✓	✓	✓	✓	✓	✗	✗	✓

Figure 8. Stock price directional predictions for ticker-day dimension



Member	Ticker	Link
yCharts	M	Link
yCharts	GOOG	Link
harmongreg	AAPL	Link
BlueFinder	AAPL	Link
MOFinancial	STX	Link
1nvestor	MBI	Link
Idrogen	LKND	Link
FinanceTrends	GLD	Link
Stocktwits	IBM	Link

Figure 9. Search by ticker and/or author

Based on date: 3/4/2013

Stock price directional Predictions for 10 days. Based on author, community and market information.

Author	Ticker	Bullish	Bearish	Sentiment	Prediction is likely to be									
					1	2	3	4	5	6	7	8	9	10
yChart	M	24	3	1.833	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗

Figure 10. Stock price directional predictions for author-ticker-day dimension

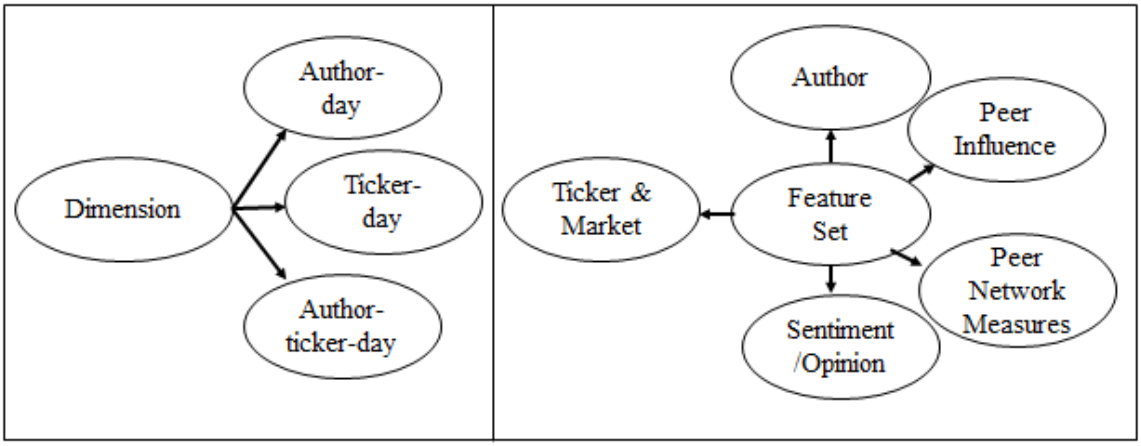


Figure 11. Dimension and feature sets

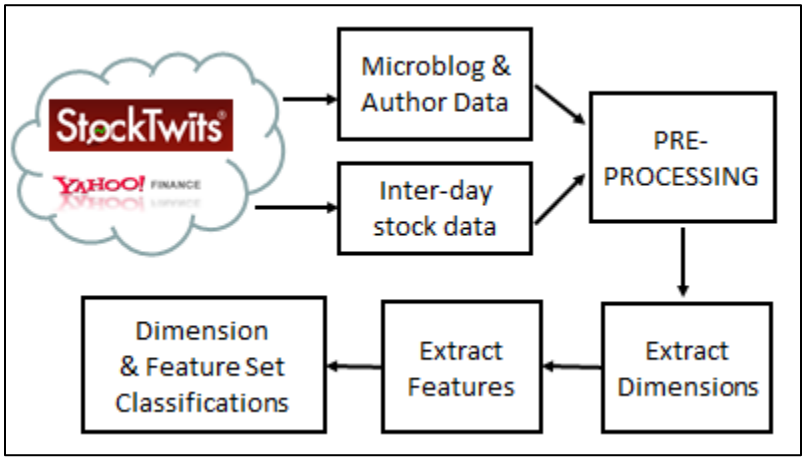


Figure 12. Process flow

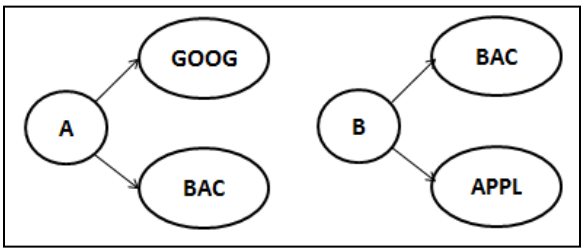


Figure 13. Authors with tickers in their microblogs

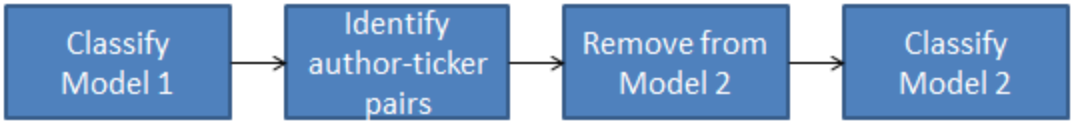


Figure 14. Process flow of adaptive learning algorithm

CHAPTER 5

EXPLORATORY STUDIES

Introduction

The nascent channel of stock microblogging offers a new perspective into the activities of irrational investors (De Long et al., 1990) that scholars have yet to fully comprehend. It offers a fresh flood of user-generated content (UGC) (Andreas & Michael, 2010) that revolve around knowledge contribution of stock market investing information. Furthermore, stock microblogging's features of succinctness, real-time information, and high volume tweets parallel market trends and thus may have the ability to capture valuable market information (Oh & Sheng, 2011; Sprenger & Welpe, 2010). This new channel presents a variety of research opportunities from content of the microblog to the network structure of co-tweeted tickers; and from manual sentiment annotation to imbalance classification of microblog features. These exploratory studies may assist in uncovering new features as well as new outcome indicators with the objective of improving predictive accuracy of these classification or explanatory models. In this chapter I present other exploratory studies that I plan to develop further in future research.

There are two exploratory topics in this chapter. The first covers network structure of co-tweeted tickers and the second examines manual annotation of sentiment via crowd sourcing.

The first topic, exploring network structure of co-tweeted tickers, involves extracting relationships among tickers that are co-cited in the same microblogs. Such relationships among tickers may lead to inferences from investors in relation to groups of tickers that move together in stock market dynamics.

The second topic, manual annotation of sentiment via crowd sourcing, examines approaches and best practices in crowd sourcing involving nonexperts. Manual labeling is important as it provides the base for automatic labeling of investor sentiment or opinion. Without sentiment it is impossible to determine outcomes such as predictive accuracy of stock price movement or information quality.

Network Structure of Co-tweeted Tickers

Introduction

Virtual Investing Community as a Social Network

A VIC is essentially a social network, comprised of many participants (hereafter authors) conversing in discussions about stock market topics. With this in mind, I argue that examining VICs through a social network perspective is not only appropriate but also novel. Even though social network analysis is currently a popular research approach (Oinas-Kukkonen et al., 2010), I find a lack of research initiatives investigating the impact of network structure surrounding stock tickers in the context of VIC. This is interesting because upon closer examination of these online conversations, one can imply

the existence of latent relationships among tickers co-tweeted in the same tweet or by the same author. The intuition is that authors tend to present a coherent message when posting by categorizing tickers of the same characteristics in the same tweet. For example, tickers that are mentioned together are trending either in the same or in the opposite stock price directions, tickers that are in the same industry, be it competitors or vendors. These relationships may present valuable investing information that correlate with present or future financial market outcomes. A few examples of co-tweeted ticker discussions from Stocktwits are presented in Table 22.

Research Questions

In this study I examine network structures of co-tweeted stock market tickers mentioned in online conversations as well as by each author. I put forward the following research questions:

1. What is the best approach to extract network structure from tickers co-tweeted in a stock microblog or by an author?
2. Is there any relationship between these network structures and financial market indicators?

Proposed Study and Contributions

I propose to use a social network approach to understand the latent relationships among co-tweeted tickers in stock microblogging discussions (hereafter tweet) with financial market outcomes. I examine 20,479 stock tweets downloaded from Stocktwits over a period of 9 days in May 2010. From these tweets, two types of networks are

constructed: a same-tweet network and a same-author network. From these networks I generate network measures of centrality (betweenness, closeness and degree) and prestige (page rank) for each ticker in the networks. I then perform OLS linear regression analysis regressing these measures with stock market indicators of trade volume, volatility, beta (risk) and return. Although the preliminary results show a low adjusted R-squared I seek to find evidence of relationships between network structures of co-tweeted tickers with market indicators.

Contributions of this study include proposing an approach to extract investing information from network of relationships among stock tickers and providing evidence in support of the value of UGC. I seek to contribute to investors in coming up with measures to better assist in investing decision making. I also provide new measures to VIC owners to design better features for their communities. For researchers, we I propose a set of measures as well as approach to extract value from UGC in online stock microblogging communities.

Relating Social Network Structure to Performance Outcomes

Social network analysis is built upon the knowledge that relationships between network nodes or actors can be represented by a graph. The graph's nodes represent the actors while the graph's edges or links represent social interactions between actors. Thereafter graph theory (Wasserman & Faust, 1994) can be applied to discover relationships among actors in the network.

Many scholars have concluded that network structures do influence economic outcomes. Granovetter (1985) is among the first to conclude that most economic behavior

in society is embedded in social relation networks. Studies have discovered relationships between network structure and biotechnology startups relationships with other companies (Walker et al., 1997), firm acquisition and cost of capital (Uzzi, 1999), and formation of new alliances (Gulati & Gargiulo, 1990).

A closely related area of research is the study of co-citation networks. One literature stream investigated company co-citations in news articles and another on author co-citations in scholarly communities. Bernstein et al. (2002) identified companies from Yahoo! News and constructed an undirected and unweighted intercompany network and reported that top 30-ranked companies in the computer industry are also Fortune 1000 companies. Ma et al. (2009) examined company co-citations in Yahoo! Finance news articles in predicting the direction and strength of revenue relations between those companies. The authors extracted dyad degree-based, node degree-based and node centrality measures from company co-citations. They concluded that more global measures such as node degree and node centrality based measures are better predictors than dyad degree-based measures. Another interesting study (Ma et al., 2011) further extended the same approach by relating company co-citations in online news to infer competitor relationships between them.

The approach of this study is largely influenced by literature on co-authorship networks where scholars focus on the structure of scientific collaborations. One such that is closely related is Liu et al. (2005) who uses social network analysis of centrality and prestige measures to examine status and influence of co-author networks in Advances in Digital Libraries (ADL) and the ACM Digital Libraries (DL) conferences. The measures employed in this study are degree, closeness, betweenness, eigenvector centrality and

weighted page rank. I extended the approach from Liu et al. (2005) and to explain relationships between network structures of stock ticker co-citations and financial market indicators. Other studies in author co-citation include communities such as JASIST (He & Spink, 2002), ACM SIGIR (Smeaton et al., 2002), SIGMOD (Nascimento et al., 2003) and information systems (Cunningham & Dillon, 1997).

In spite of the notable scholarly effort I argue that there exist a research gap in applying a social network approach in harvesting measures from stock market ticker co-citation network structures in identifying ticker status and influence in stock microblogging communities. I also argue that even though stock microblogging is a nascent UGC channel, it is rich in investing information and very relevant to both finance and social network research.

Research Design

In this study I explore two types of ticker-to-ticker network; one is based on dyad ticker relationship found in each tweet while another is based on dyad relationship mentioned by each author. The first (named as same-tweet network) accounts for pair(s) of ticker relationships that can be derived from each tweet where more than one ticker is mentioned. The second (named as same-author network) examines the pair(s) of ticker relationships derived from all tickers mentioned by each author. Essentially I follow Bernstein et al. (2002) by aggregating simple co-tweeted tickers (either in tweet or by the author) and drawing conclusions statistically about semantic relatedness of those tickers. Although simple co-citation of any two tickers does not necessarily mean that they are related in any general sense but “statistical aggregation allows unimportant co-

occurrences to act as noise and important relationships to act as signal" (Bernstein et al., 2002, p. 4).

Same-Tweet Network

This network is based on the tickers mentioned in each tweet. The intuition here is that there are some underlying latent relationships among tickers mentioned in the same tweet. The basic model in this network is the undirected, binary graph model with each link represents a co-tweeted relationship between any two tickers. Due to the nature of the relationship between tickers I am unable to assume direction within the constructed links. Consider two tweets with their corresponding mentioned tickers. See Figure 15.

Undirected Links

A link is created for every dyad ticker that appears in the same tweet. In Figure 15, the first tweet has a link connecting 'CMG' and 'AAPL'. In the case of the second tweet, three links are formed, 'PCLN' and 'AAPL', 'PCLN' and 'MNST', and 'AAPL' and 'MNST'. The link is undirected thus there is only one link for each possible dyad in a tweet. The drawback of this model is that it ignores tweets with single ticker since no dyad link can be constructed from a single ticker. The constructed dyad links are illustrated in Figure 16.

Weighted Undirected Links

Weights are added in two ways. First is by counting the frequency of occurrence for each link and second is by generating the exclusivity measure of each ticker in each tweet (refer later section for a detailed explanation on calculating the exclusivity

measure). With the first approach the count of each link is added, as shown in Figure 17 (A). Since each one of the links in this network only occurs once, the weight is 1 for every link. In the second approach, the link in the first tweet (Figure 16(A)) has an exclusivity value of 1 while the ones in the second tweet (Figure 16(B)) have the value of 1/2. The respective values are updated in Figure 17 (B).

For further clarification I extend this discussion with Tweet 3 which has three tickers {"PCLN", "C" and "MNST"}. Thus 3 new links -- 'PCLN' and 'MNST', 'PCLN' and 'C' and 'C' and 'MNST' are formed (see Figure 18). These new additions alter the network structure of the unweighted network by adding a new node 'C' and replicate an existing link (PCLN-MNST). And for the weighted network (1st approach – Figure 18(A)), this new link adds to the count of an existing link ('PCLN'-'MNST'), resulting in a new weight of 2. And for the weighted network (2nd approach – Figure 18(B)), the exclusivity value of this link (1) is added to the existing value (.5) giving a new weight of 1.5. The weights of the new links for 'C' are also added.

Same-Author Network

This network is based on all tickers mentioned in all tweets sent by each author. The intuition here is that there are some underlying latent relationships among tickers mentioned by the same author. Consider the tweets in Figure 19 mentioned by the author 'BOSSMONEY'.

Undirected Links

Similar to the same-tweet network, an undirected binary graph modal is constructed where each link is created for every dyad ticker that is mentioned by the same author. In Figure 20, six distinct links are created for ‘CMG’ and ‘AAPL’, ‘CMG’ and ‘MNST’, ‘CMG’ and ‘PCLN’, ‘AAPL’ and ‘PCLN’, ‘AAPL’ and ‘MNST’ and ‘PCLN’ and ‘MST’.

Weighted Undirected Links

Weights are added by counting the frequency of each link in two ways: one without grouping and another grouping by author. The intuition is that a ticker may be mentioned frequently by many authors or by just a few authors. So in our example without grouping case (Figure 20(A)), the link ‘CMG’ and ‘AAPL’ appear twice, so it has the weight of 2. However in the grouping by author case (Figure 20(B)), since there is only one author, all links have the weight of 1. Figure 20 illustrates our discussion.

Weighted Undirected Links for Multiple Authors

I further extend the discussion by adding Tweet 4 from another author – ‘justin0820’ (Figure 21). Tweet 4 has three tickers (‘CMG’, ‘AAPL’ and ‘IWM’). This will add a new node ‘IWM’ into the network and three new links – ‘CMG’ and‘IWM”, ‘CMG’ and ‘AAPL’ and ‘AAPL’ and ‘IWM’. The updated networks diagram with the inclusion of Tweet 4. See Figure 22.

In the non-grouping case, the weight for CMG-AAPL is increased by 1 and two new links are added for IWM-CMG and IWM-AAPL (Figure 23(A)). Similarly for the

grouping case, CMG-AAPL weight is increased by 1 and two new links are added (Figure 23(B)).

Single Ticker Tweets

The same-author network caters for single ticker tweets in its construction. Single ticker tweets have a significant role as they encompass about 75% of the dataset. When I include Tweet 5 from 'justin0820' into the author network the respective diagrams is updated. See Figures 24 and 25.

The clear difference between links in the same-tweet network and those in the same-author network is that the links in the former are local while the latter are global. This is because in the same-author network every ticker that is mentioned by an author is paired with every other ticker. But in the same-tweet network, only those in the same tweet are paired. Hence I observe no links between CMG and other tickers beside AAPL in the same-tweet network (Figure 18) whereas CMG is linked to every other ticker in the same-author network (Figure 20).

Calculating the Exclusivity Measure

The exclusivity measure (Liu et al., 2005) represents the degree by which both tickers in a link have an exclusive co-tweeted relation in a particular tweet. This measure gives more weight to co-tweeted links that have a lower number of other tickers in the same tweet. So for a tweet with many tickers, each individual ticker-to-ticker relationship should be weighted less. The notion is that the stronger the exclusivity value, the stronger

the underlying latent relationship. Essentially a ticker with a higher exclusivity measure is more salient towards the tweet. I define:

Exclusivity = 1 / (Total-links – 1)

From the example in Figure 26 I calculate the exclusivity value for Tweet A to be 1 (1/(2-1)) and those in Tweet B to be ½ (1/(3-1)). Thus the 'GOOG-AAPL' link in Tweet A is more exclusive than 'C-INTL' link in Tweet B (see Figure 26). Thereafter, at the global level, I sum all exclusivity measures for occurrence of the link in every tweet where the link is found. Thus following the example above, I sum all exclusivity scores for GOOG-AAPL that occurs in every tweet.

Ticker Co-tweeted Measures

Centrality Measures (Degree, Closeness, Betweenness)

In this study I use three measures of centrality: degree, closeness and betweenness centralities (Freeman, 1978) to determine the relative importance of a node in a network. All social network measures are generated using social network analysis software JUNG (O'Madadhain et al., 2006).

Degree Centrality

Degree centrality is the measure of the total number of links that are connected to a particular node. It represents a basic notion of centrality by measuring how many connections each node or author has to their immediate neighbors in the network. A disadvantage with degree centrality is that it precludes any information about the node's link to other nodes beyond its immediate neighbors. In short while it considers only the

local structure it ignores the global structure of the whole network. In addition, degree centrality for the same-tweet network is different from the same-author network. Considering the same two examples from the previous section I generate degree centrality for each ticker (see Table 23).

Note that in the same-tweet network AAPL, MNST and PCLN have the highest degree centrality (3) being the nodes that connect to three other nodes. In the same-author example, CMG and AAPL have the highest degree (5) each with five connections.

Closeness Centrality

Closeness centrality is the second type of centrality measure that extends degree centrality by determining how close a node is to all other nodes in the network, beyond just adjacent neighbors. Essentially it looks at the shortest path distances of a node to all other nodes. Thus a node is more central the lower its total distance is to all other nodes representing the lower amount of time needed to spread of information. I generated closeness centrality values for each ticker (see Table 24).

Similar to degree centrality, in the same-tweet network AAPL, MNST and PCLN have the highest closeness centrality (.8) being the nodes that are closest to all other nodes in the network. In the same-author example, CMG and AAPL have the highest closeness score (1).

Betweenness Centrality

Betweenness centrality is the measure of how likely a particular node is on the shortest path between any pair of randomly chosen nodes. Such nodes act as bridges and

are assumed to be highly central since they are in the path of information flow in the network. Betweenness centrality is also defined to be a reliable measure of structural holes (Burt, 1992). I generated the following betweenness centrality for each ticker (see Table 25).

Note that in the same-tweet network AAPL has the highest betweenness centrality measure as it is the only one in the shortest path between any pair of nodes in the network. But in the same-author network, both AAPL and CMG are in the shortest path.

Prestige Measures (PageRank)

PageRank (PR) is a centrality algorithm made popular by Google. It models inherited or transferred status (Liu et al., 2005) by assigning a numerical weighting to each element of a hyperlinked set of documents with the purpose of measuring its importance within the set (Wikipedia, 2012). I generate PR scores for 4 sets: same-tweet network (frequency), same-tweet network (exclusivity), same-author network (no grouping) and same-author network (group by author). The values are shown in Table 26.

I find that for the same-tweet network, MNST and PCLN have the highest PR scores due to their weights from both frequency (9.82) and exclusivity (6.16). In the same-author network, AAPL and CMG ranked highest in the group by author weight while only AAPL has the highest PR score in the nongrouping weight

Analysis and Results

The dataset in this study consists of 20,479 tweets downloaded from Stocktwits, between May 11, 2010 and May 19, 2010. This dataset pertains to 2,075 authors and 995

tickers. These tweets are pre-processed to extract ticker, date and time. From this dataset I constructed a same-tweet network of 12,184 links and later analyzed using the social network software JUNG (O'Madadhain et al., 2006) producing 3,267 ticker-per-day data points. Simultaneously, a same-author network of 10,609 links is created which is then analyzed and produced 785 ticker-per-day records. A process flow diagram is shown in Figure 27.

Constructing the Same-Tweet Network

Links are created between each co-tweeted ticker dyad mentioned in each tweet for each day in the dataset. Top 10 same ticker links are shown in Figure 27. Tweets for single tickers are ignored as links are unable to be constructed from single tickers. This results in an undirected same-tweet network. I then include weight by two approaches: 1) Adding count of frequency of a link and 2) generating exclusivity measure for each.

Constructing the Same-Author Network

Links are created between author and each ticker mentioned by the author for each day in the dataset. Top 10 author-ticker links are shown in Figure 28. I then transform the network from a 2-mode network (author-to-ticker) to a 1-mode network (ticker-to-ticker) (de Noy et al., 2005). This creates undirected dyad relationships among tickers by the same authors. I then aggregate these links generating frequency counts for each link as a measure of weight. Descriptive statistics for both same-tweet (Figure 29) and same-author network (Figure 30) are provided. In addition, top 10 tickers ranked by all measures are also available (Figure 31).

Variables

Independent variables consist of network measures of centrality: degree, betweenness, closeness and Page Rank. Dependent variables are market indicators: % return, Garman volatility, Parkinson volatility, beta and trading volume.

OLS Regression for Same-Tweet Network

Preliminary OLS Regression Results regressing DV (market indicators) on network measures from same-tweet co-tweeted network. Market indicators are same day measures of volatility: Garman volatility (Garman & Klass, 1980) and Parkinson volatility (Parkinson, 1980), trade volume and return. The four measures of IV are betweenness centrality, closeness centrality, degree centrality and page rank. But page rank is ignored in the model due to SPSS issue. See Table 32.

OLS Regression for Same-Author Network

Preliminary OLS Regression Results regressing DV (market indicators) on network measures from same-author co-tweeted network. Both IVs and DVs are the same as used in same-tweet network analysis. See Table 33.

Discussion and Conclusion

Referring to the Table 32 and 33 I conclude that despite the low adjusted R-squared, I do observe evidence of relationships between network measures of co-tweeted tickers and market indicators. However the relationships for nodes generated from the same-tweet network are more salient than those generated from the same-author network.

Intuitively this implies that dyad tickers that are co-tweeted in the same stock tweets do have latent relationships that correlate with financial outcomes. I observed that a ticker that is high in betweenness centrality, indicating a node that is often in the shortest path between any two random nodes in the network, tend to be more volatile, gives higher return, higher beta but with lower trading volume. Thus tickers such as GLD (betweenness=41618), APPL (betweenness=35591) and C (betweeness=33136) that are frequently co-tweeted with many other tickers are in this group. Conversely, a ticker that is high in closeness centrality tends to correlate with lower volatility, lower return and lower beta but higher trading volume. Similarly, a ticker that is high in degree centrality, indicating a high number of connections to adjacent neighbors, tend to correlate with lower volatility, lower return and lower beta but higher trading volume. Thus tickers such as TIVO (degree=85), BIDU (degree=76) and NFLX (degree=121) are in this group. I also observed that a number of tickers (i.e. BP, TER, NFLX, IWM) in the top PR ranking are not present in the betweenness and degree rankings. The difference is due to the presence of weights in PR algorithms as the other centrality measures do not consider weight of each node. This notion of weight refers to the exclusivity and frequency measure of each ticker in a tweet. Thus, while other centrality algorithms disregard the conversation intensity surrounding a ticker, the PR accounts for this element and produces a more realistic ranking.

In general, tickers that are frequently co-tweeted with other tickers are more centrally located in the network indicating that these tickers are more discussed in conversations and tend to have a more salient correlation with market outcomes. However, those tickers that are more connected to adjacent neighbors correlate more with

higher trading volume but with lower returns. This seems to be related to the financial phenomenon of popular stock (Bauman, 1965) which is associated with lower returns. On the other hand, tickers that act as bridges to all tickers in the network tend to correlate with lower trading volume but higher returns. Such is related to the concept of structural holes (Burt, 1992).

Contributions of this study include proposing an approach to extract investing information (metrics) from network of relationships among stock tickers and providing evidence in support of the value of UGC. I seek to contribute to investors in coming up with measures to better assist in investing decision making. I also provide new measures to VIC owners to design better features for their communities. For researchers I propose a set of measures as well as approach to extract value from UGC in online stock microblogging communities. I seek to improve this study by achieving a higher relevant result to support the findings.

Sentiment Annotation Via Crowd Sourcing

Introduction

The evaluation of IR systems uses ground truth (gold sets), which is commonly obtained from experts who manually judge relevance of document-class pair. But experts are both costly and time-consuming. Fortunately, such tasks can be outsourced as micro-tasks on the web to anonymous web users or non-expert annotators via providers such as Amazon Mechanical Turk (MTurk) or CrowdFlower (Vuurens et al., 2011). In fact, MTurk has attracted increasing attention in practitioner and academic research as a convenient, inexpensive, and efficient platform for crowdsourcing (Tang & Lease, 2011).

Nonetheless, crowdsourcing is not the panacea to manual annotation and has its disadvantages due to the low quality of unknown work force specifically from malicious intent and sloppy work. In fact some claim that up to 90% of data from crowdsourcing are worthless (Stone et al., 2011). One common solution is to rely on repeated labeling (Sheng et al., 2008) and simple Majority Voting (MV) (Sheng et al., 2008; Snow et al., 2008) to identify the correct labels where annotation that receives the maximum number of votes is the final aggregated label. But this technique not only increases cost but is less useful when the majority of labels come from noisy or low quality workers. Hence having the ability to filter good workers from inferior ones may greatly improve the quality of judgments submitted. In fact, Wang et al. (2010) (another Panos and Foster's paper) found that by assessing worker quality and removing inferior workers, the overall quality greatly improved. To help with this scholars have identified different categories of crowdsourcing workers. For example, Vuurens et al. (2011) concluded 4 classifications of workers (Vuurens et al., 2011): proper worker (ethical worker), random and semi-random spammer (unethical worker: I merge both into one classification), uniform spammer (unethical worker) and sloppy worker.

The initial approach in filtering quality workers is through qualification tests (Stone et al., 2011) where only those who are qualified (based on threshold of accuracy score e.g. >=80%) are allowed to perform the tasks. This qualification tests can be treated as an initial training period (Le et al., 2010) where workers are notified of mistakes and given opportunities to retake the tests.

As worker quality varies from time to time, evaluation needs to be performed on a continuous basis, preferably after each batch of labels. The "honey pot" or "trap

questions" approach is a good way to do this. "Honey pots" (Stone et al., 2011) or "trap questions" (Tai et al., 2011; Zhu & Carterette, 2010) refer to insertion of labeled examples (ground truths) among the unlabeled examples to monitor worker quality on-going. The accuracy of each worker can be generated from worker labels of these ground truths and appropriate actions are taken to remove or improve worker quality.

Both qualification tests and "honey pots/trap questions" approaches ensure that only qualified workers are initially accepted and continuously engaged.

Another approach to estimate worker quality is using machine learning. Although supervised consensus labeling in monitoring worker quality is best, it is not feasible in handling large volume of tweets. A more sensible approach is introduced by Tang & Lease (2011) by using a semisupervised approach. This technique involves generating a classification model from a small dataset of labeled examples (ground truths) to generate "soft labels" for unlabeled examples. These soft labels are then used to evaluate worker quality. In this study I propose two types of classifiers for this task: first is the Naïve Bayes classifier on bag-of-words feature dataset and second is a lexical classifier on lexicons of bullish and bearish word lists.

Research Design

The process of annotating is performed in small batches (Soleymani & Larson, 2010) so that worker quality can be continuously updated and monitored. In addition, this approach reduces worker fatigue. Each tweet is to be labeled by three workers and a majority vote (MV) is ascertained from the three labels (McCreadie et al., 2010). A

fourth label by an expert is required if MV is not available for any tweet. The details of the labeling procedure are as per Figure 28.

Pre-test

I qualify workers with a pretest (20 tweets with ground truths to label (Le et al., 2010)) and only accept those who qualified with an acceptable quality threshold (e.g. >= 80% score). Those who do not qualify may take future tests. Feedback is provided to workers to improve their quality. This pre-test needs to be done periodically to obtain new workers as older ones dropped out due to lower quality or lack of interest/motivation.

Label Batch

Each batch consists of one to two periods of tweets (N is about 7000). Each tweet is to be labeled by three workers in order to obtain Majority Vote (Sheng et al., 2008; Snow et al., 2008; Yang et al., 2010).

Honey Pot Test

The literature suggests insertion of “honey pots/trap questions” into each batch to continuously monitor worker quality (Tai et al., 2011; Zhu & Carterette, 2010). Since MTurk does not provide such features, I perform this step after each batch is completed where N labels are randomly collected from each worker and evaluated against ground truths. Those workers who do not qualify are removed from the worker list and their labels are reassigned to other workers.

Majority Vote Processing

I obtain Majority Vote for each tweet based on the three labels from MTurk workers. A fourth label is needed from experts for those tweets without majority vote. All tweets should have majority vote (e.g., Batch 1 nonmajority tweet is 4.7% while Batch 2 non-majority tweets is 5.8%).

To ascertain some level of confidence with these MV labels, I random sample N number of tweets to be manually labeled by experts. A pairwise Pearson correlation is then conducted to gauge how close crowd MV labels are with expert labels. (e.g., Batch 1 Pearson correlation =.908 while Batch 2 Pearson correlation =.89).

Evaluate

I evaluate the crowd labels by comparing the quality with those from experts via bi-variate correlation analysis.

Data and Results

Data

The stock tweets for this study are downloaded from Stocktwits.com. This dataset covers the period dated from 3/2/2012 to 3/15/2012 totaling 38,202 tweets. I divided the sample into eight smaller batches for crowd annotation. Detailed information for each batch is in Table 34. Note that cost (denoted by *) consists of cost of labels (2 cents per label), fees to MTurk (25% of total cost), cost of relabeling (about 10% of labels). And estimated cost for experts (denoted by **) is $15 per hour for 100 labels for three experts.

Process

I conducted three pretests to identify/locate qualified workers. Each pretest consists of 20 tweets to be labeled by 100 workers = 2000 labels. Three hundred and thirty-two workers attempted the pretests, out of which only 135 qualified based on having at least 15 labels at 80% accuracy threshold.

For each batch, I conducted a honey-pot test whereby a random sample of five labels per worker is collected and manually labeled by an expert to determine ground truths. I then compare worker labels with ground truths and generate accuracy score per worker. The average accuracy score and worker count per batch is listed in Figure 35. There are in total 71 distinct workers who label this dataset.

I obtained majority vote for each batch and calculated the percentage of tweets without MV. The average of no MV is 6.74%. These tweets are then reassigned to an expert to get majority vote. See Figure 36.

To evaluate the quality of crowd labels, I perform bi-variate correlation analysis between 100 random sample labels from the MV and corresponding labels from an expert to gauge how similar are the labeling qualities. I compared Pearson and Spearman correlation coefficients and found satisfactorily results to support the closeness of labeling quality between workers and expert. Due to time constraint and the positive results, only the first four batches were processed. See Figure 37.

Discussion and Conclusion

Although this is just an exploratory study, the preliminary findings conclude that the quality of labels from crowd annotation is as good as those from the experts. This is

clearly a positive sign as crowd labels offered a lower cost and high availability option to researchers and practitioners. In this study I propose a framework in crowd annotation of stock tweet labels extending past scholarly studies. I discuss the findings below.

The three vital parts of the proposed framework all focus on ensuring quality workers and quality results. The three parts are: 1) pretest, 2) honey-pot test and 3) majority vote. It is key to first start with the right group of workers and then to continuously improve their performance results and motivation by processing in smaller batches. Workers are motivated by prompt payments as well as acknowledgement of their effort. From this initial result, I conclude that the proposed framework is able to develop a quality team of annotators who are proficient and efficient.

The cost of using crowd annotation is much lower. From this study, I note that the cost is about 5 times more for expert labelers at $15 per expert for an hour. The cost of labeling 38202 tweets for crowd is $3,397.00 while the estimated cost for experts is at $17,190.00 (Table 34). This is a major factor in labeling due to the high volume of labels that needed to be processed. High availability is another factor that is encouraging for crowd annotation. Crowd sourcing channels such as MTurk provide thousands of available workers around the clock. Furthermore I am able to perform basic filters to ensure workers are conversant in English and have high past success rates. High availability leads to fast turnaround for processing each batch. This the third advantage of crowd annotation.

Crowd sourcing does have its disadvantages and one such is the lack of communication opportunities. Naturally, crowd sourcing channels try to minimize communications between providers and workers to prevent loss of business. The work

around is to plan for an additional point of contact via requiring workers to register themselves elsewhere. The more concerning factor is the quality of workers in crowd annotation. The best practice is to continuously improve quality by evaluation and feedback. With crowd annotation time and investment are needed to develop a team of workers who are eager and knowledgeable in a particular domain such as stock tweet labeling.

One future extension is to perform more rigorous evaluation with expert labels as a comparison to crowd annotation. This would reinforce the findings and support the viability of the proposed framework.

Table 22. Examples of Stocktwits microblog posts (tweets)

Tickers	**Microblog/Tweet**	**Implied ticker relationship**
$SAM, $ABV, $TAP	Beverages-Brewers doing well today... $SAM all time high on guidance, $ABV all time high on upgrade, $TAP 52 week high	In the same industry and market trend
$VZ, $AAPL	Apple Will Sell 9 Million Verizon iPhones Next Year, Says Gene Munster, But He Hints It Could Be Double That $VZ $AAPL	Vendor relationship

Table 23. Degree centrality measures

	AAPL	CMG	MNST	PCLN	C	IWM	AMZN
Same-tweet	3	1	3	3	2	-	-
Same-author	5	5	3	3	-	3	3

Table 24. Closeness centrality measures

	AAPL	CMG	MNST	PCLN	C	IWM	AMZN
Same-tweet	.8	.5	.8	.8	.571	-	-
Same-author	1	1	.71	.71	-	.71	.71

Table 25. Betweenness centrality measures

	AAPL	CMG	MNST	PCLN	C	IWM	AMZN
Same-tweet	3	0	1	1	0	-	-
Same-author	2	2	0	0	-	0	0

Table 26. PageRank measures

	AAPL	CMG	MNST	PCLN	C	IWM	AMZN
Same-tweet (frequency)	6.59	2.01	9.82	9.82	5.98	-	-
Same-tweet (exclusivity)	3.91	1.9	6.16	6.16	2.99	-	-
Same-author (group by author)	2.51	2.51	1.49	1.49	-	1.49	1.49
Same-author (without grouping)	3.08	2.66	1.85	1.85	-	1.2	1.2

Table 27. Top 10 ticker-ticker links

Ticker	**Ticker**	**Links**
AAPL	GOOG	68
V	MA	50
SAP	SY	48
JPM	GS	42
JPM	C	38
C	JPM	38
C	GS	36
GS	MS	34
BP	RIG	33
GS	BAC	32

Table 28. Top 10 author-ticker links

Author	Ticker	Links
BlueFielder	AAPL	90
TraderFlorida	AAPL	42
theback9	GS	41
Wfctrader	V	40
Paulwoll	GLD	35
Demonicshiksa	BP	32
Wsmco	UCO	31
Valeriobrl	GOLD	29
Demonicshiksa	AAPL	29
Paulwoll	GMCR	27

Table 29. Descriptive statistics for same-tweet network

	N	Minimum	Maximum	Mean	Std. Deviation
betweenness	3267	0	38509.7	557.7677	2245.324
closeness	3267	0.108521	1	0.537344	0.344857
degree	3267	1	39	3.104071	3.377396
pagerank	3267	1.57E-33	5E+175	2.2E+173	8.7E+144
volume	3267	0	9.44E+08	17875374	67409495
g_vola	3267	0	0.396996	0.001149	0.007528
p_vola	3267	0	0.638393	0.001286	0.012198
beta	2909	-2.65	7.23	1.268848	0.990855
ret	3267	-41.0673	206.061	-0.45339	4.870255

Table 30. Descriptive statistics for same-author network

	N	Minimum	Maximum	Mean	Std. Deviation
betweenness	785	0	8301.38	214.4667	706.249
closeness	785	0	1	0.431936	0.20275
degree	785	1	121	8.24586	13.88443
volume	785	0	9.44E+08	24878897	76396083
g_vola	785	0	0.200982	0.001537	0.008098
p_vola	785	0	0.190784	0.001611	0.008795
beta	677	-0.16	4.87	1.244978	0.937066
ret	785	-41.0673	49.9099	-0.26655	4.303876

Table 31. Ticker ranked by measures of betweenness, closeness, degree and PageRank.

Between			Closeness			Degree			PR		
GLD	19-May-10	38509.7	CHL	14-May-10	1	AAPL	17-May-10	39	BP	17-May-10	4.91E+175
AAPL	17-May-10	35591.3	DCTH	18-May-10	1	GLD	19-May-10	38	AAPL	17-May-10	3.02E+175
C	14-May-10	33136.2	GOLD	18-May-10	1	AAPL	14-May-10	38	GLD	17-May-10	2.57E+175
AAPL	18-May-10	32602.6	AKNS	18-May-10	1	AAPL	18-May-10	36	C	17-May-10	1.87E+175
AAPL	19-May-10	28132.7	ANV	11-May-10	1	GLD	18-May-10	36	CL	17-May-10	1.60E+175
GLD	18-May-10	27270	BMY	12-May-10	1	AAPL	19-May-10	35	TER	17-May-10	1.5887E+175
AAPL	14-May-10	24737.6	CTL	13-May-10	1	GLD	14-May-10	33	NFLX	17-May-10	1.5452E+175
GLD	14-May-10	19921.5	UNH	17-May-10	1	GS	17-May-10	32	GS	17-May-10	1.4682E+175
GS	17-May-10	18980.2	COCO	19-May-10	1	C	14-May-10	31	S	17-May-10	1.4243E+175
GS	18-May-10	17913.8	IAG	11-May-10	1	AAPL	12-May-10	30	IWM	17-May-10	1.4139E+175

Table 32. Standardized coefficients for OLS regression on market indicators (Same tweet networks)

IV	**DV**				
	Garman Volatility	**Parkinson Volatility**	**Volume**	**% Return**	**Beta**
Betweenness	.116(0)***	.084(0)**	-.051(999)	.094(.000)***	.101(.000)***
Closeness	-.037(0)**	-.029(.001)	.04(3441657)**	-.005(.257)	-.052(.055)***
Degree	-.113(0)***	-.079(0)**	.306(677204)****	-.107(.051)***	-.076(.011)**
Page Rank	NA	NA	NA	NA	NA
Adj R-squared	.003	.001	.064	.002	.004
F	3.59	1.95	57.12	2.36	3.91
N	3266	3266	3266	3266	2908

Robust standard errors in parenthesis. *$p<.10$, ** $p < .05$, *** $p < .01$, **** $p < .001$

+ Logged

Table 33. Standardized coefficients for OLS regression on market indicators (Same author networks)

	Garman Volatility	**Parkinson Volatility**	**Volume**	**% Return**	**Beta**
Betweenness	.162(0)**	.257(0)***	-.068(8215)	-.13(0)*	-.006(0)
Closeness	-.031(.001)	-.023(.002)	-.023(1E+007)	.025(.758)	-.113(.173)***
Degree	-.574(.001)	-.9(.001)	2.077(7281597)	-.756(.416)	.643(.095)
Page Rank	NA	NA	NA	NA	NA
Adj R-squared	.002	.015	.038	.011	.01
F	1.45	4.0	8.82	3.2	2.68
N	784	784	784	784	676

Robust standard errors in parenthesis. *$p<.10$, ** $p < .05$, *** $p < .01$, **** $p < .001$

+ Logged

Table 34. Batch details

Batch	Period	N tweets	N Labels	Date	Cost	Estimated cost for expert labelers
1	94-95	7032	NA	3/12-3/13	577.4	
2	96-97	9087	NA	3/14-3/15	1116.8	
3	89	4164	12492	2-Mar	322.71	
4	90	3299	9897	5-Mar	255.67	
5	91	4325	12975	6-Mar	335.18	
6	92	3637	10911	7-Mar	281.86	
7	93	3428	10284	8-Mar	265.67	
8	98	3230	9690	16-Mar	242.25	
TOTAL		38202			3397.54*	17,190**

Table 35. Average worker accuracy per batch

Batch	Worker Count	Average Accuracy
1	27	0.88
2	37	0.859
3	28	0.853
4	30	0.82
5	33	0.842
6	31	0.847
7	31	0.89
8	30	0.826

Table 36. Majority vote proportion per batch

Batch	Total	MV	No MV	% No MV
1	7032	6698	334	4.75%
2	9087	8559	528	5.81%
3	4164	3882	282	6.77%
4	3299	3124	175	5.30%
5	4325	4037	288	6.66%
6	3637	3348	289	7.95%
7	3428	3195	233	6.80%
8	3230	2911	319	9.88%
Avg				6.74%

Table 37. Correlation coefficients for batches

Batch	Pearson	Spearman
1	0.908	0.889
2	0.89	0.88
3	0.883	0.887
4	0.845	0.845

Figure 15. Examples of dyad tickers mentioned in tweets

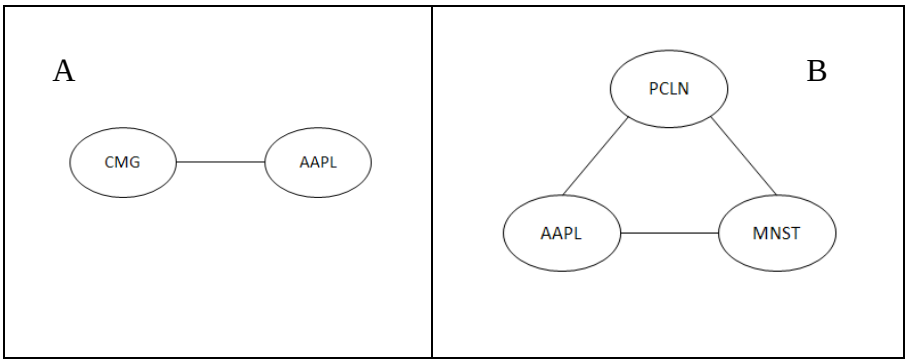


Figure 16. Undirected links

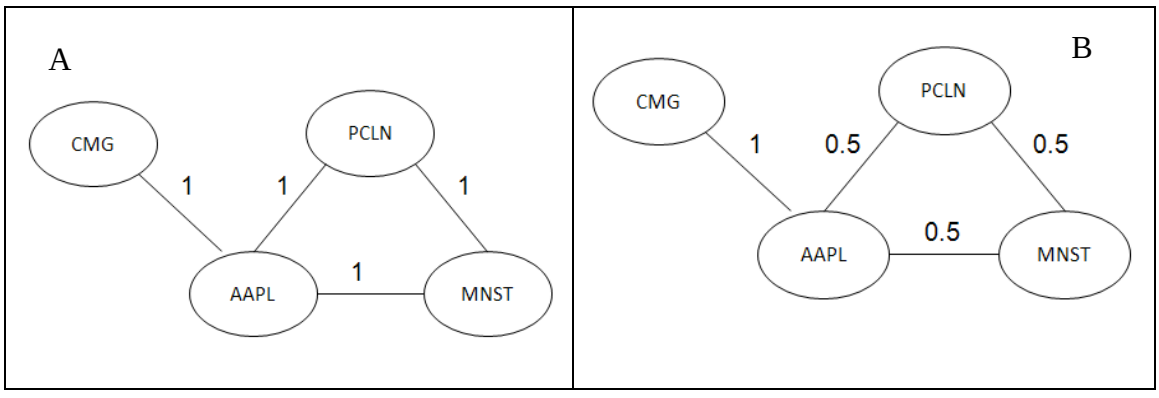


Figure 17. Weighted undirected links

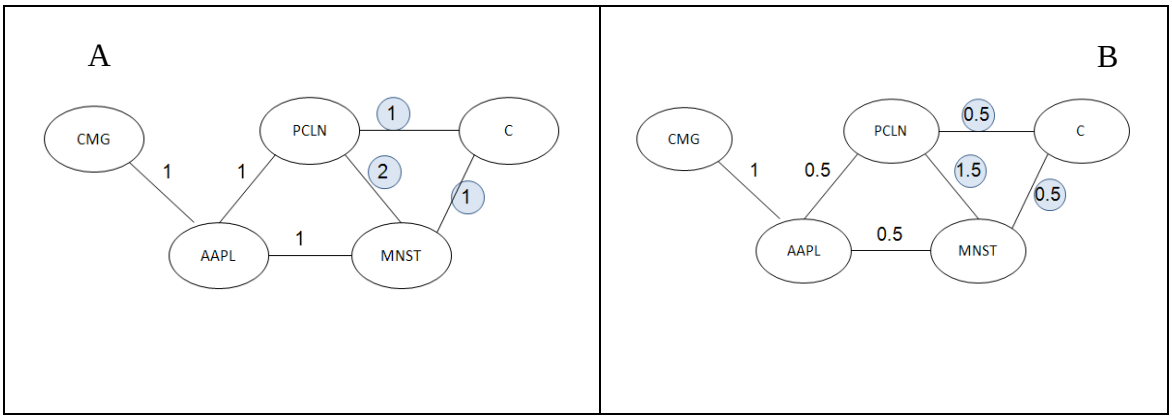


Figure 18. Weighted undirected links extension

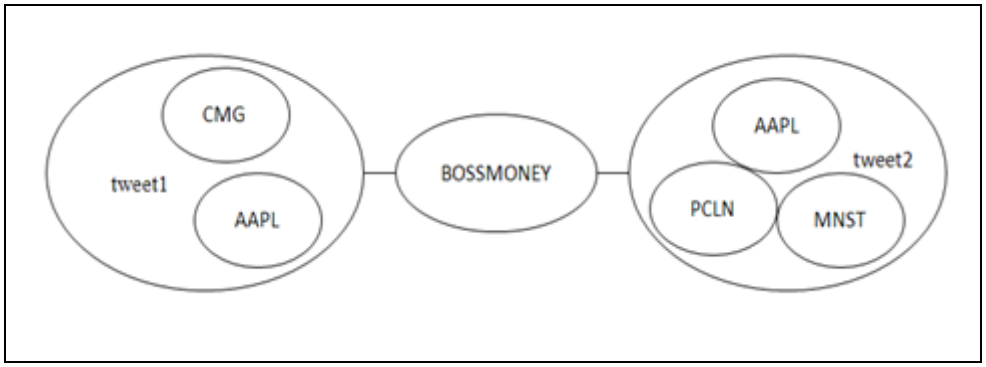


Figure 19. Same author network

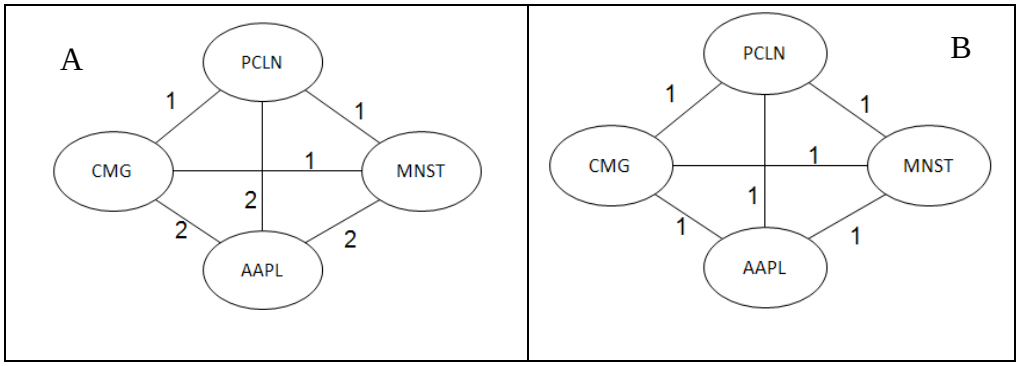


Figure 20. Weighted undirected links.

Figure 21. Tweet 4

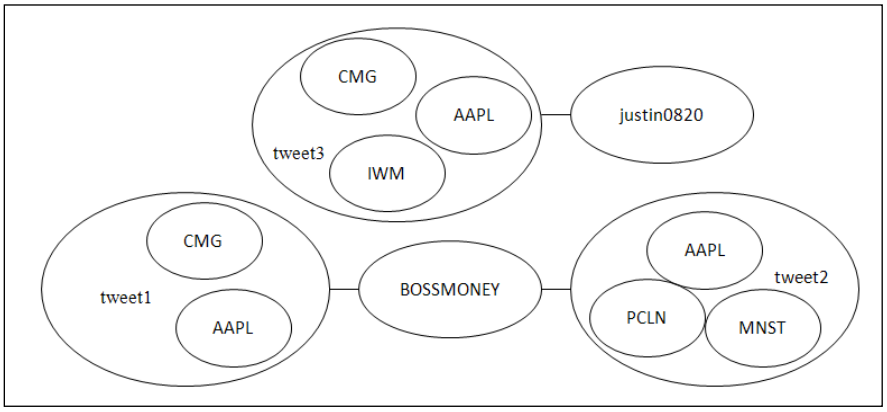


Figure 22. Network structure of weighted undirected links with multiple authors

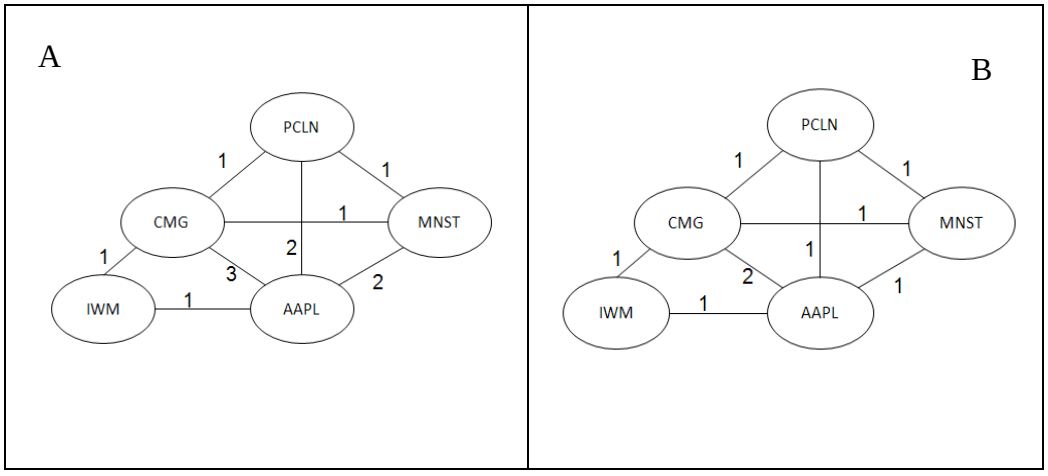


Figure 23. Without and with grouping structure.

Figure 24. Tweet 5.

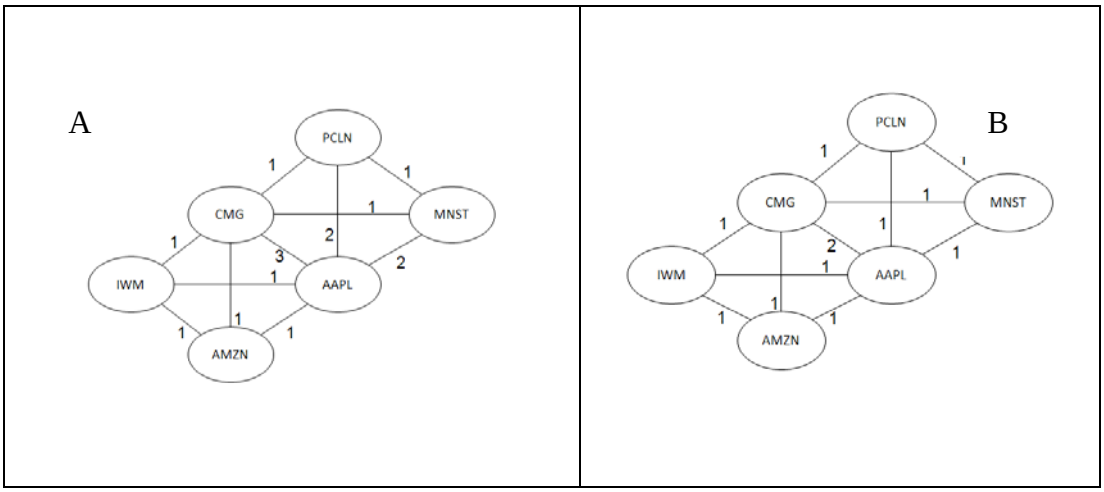


Figure 25. Without and with grouping structure.

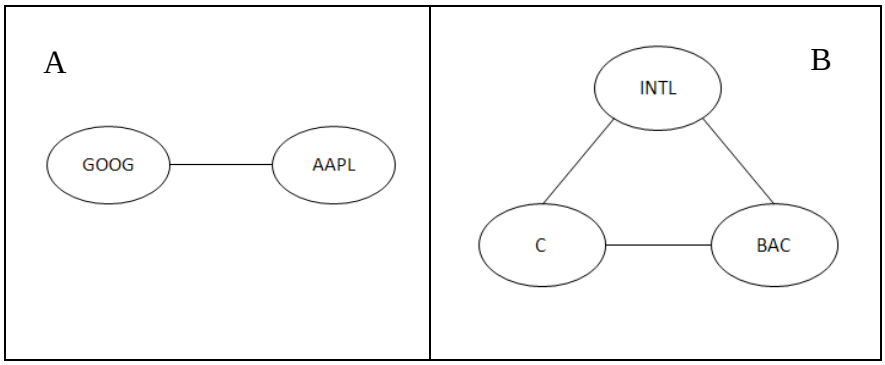


Figure 26. Exclusivity measure

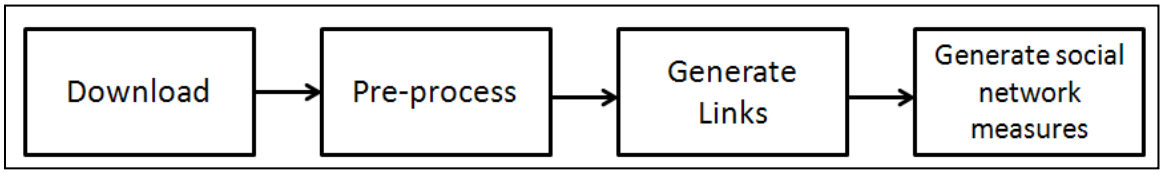


Figure 27. Process flow of network measure generation.

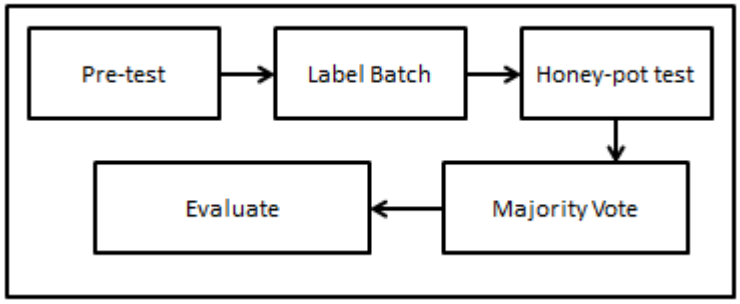


Figure 28. Workflow of labeling process

APPENDIX

SUPPLEMENTARY INFORMATION

Stocktwits Profile Page

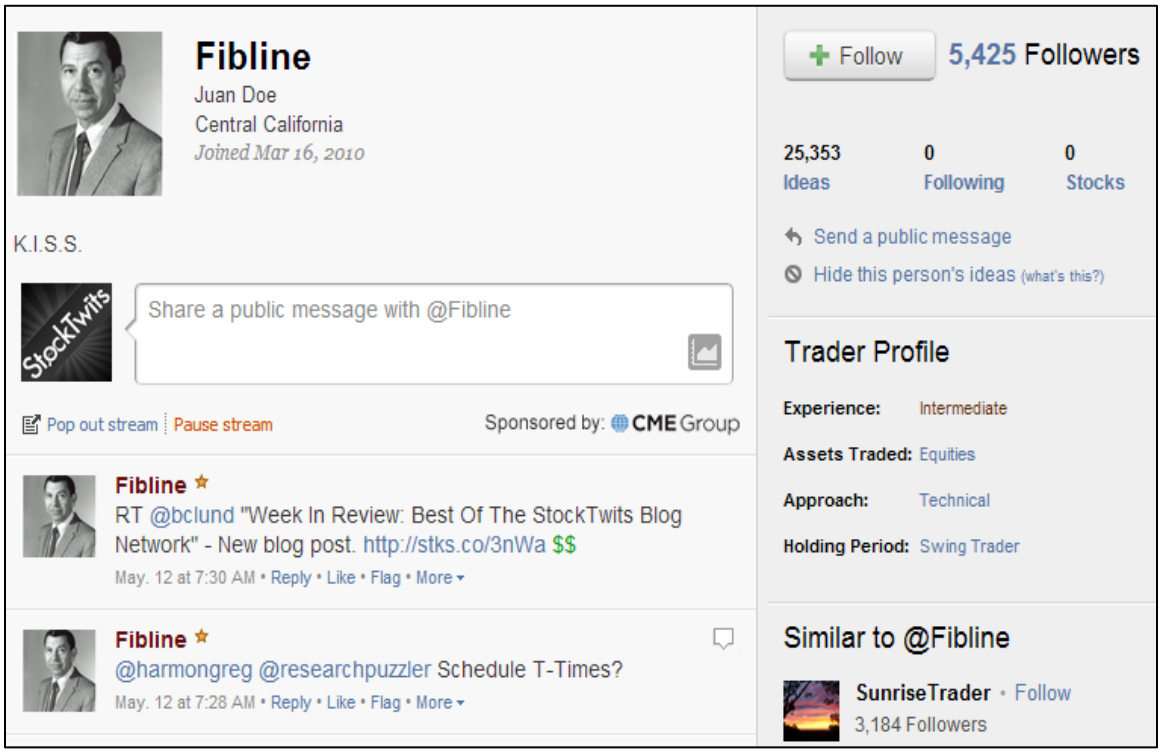


Figure 29. Stocktwit profile page.

Examples for Measures of Normalized Influence by Peer Outdegree and Normalized Influence by Unique Peer

An example of *Normalized influence by peer outdegree* (NIPO) and *Normalized influence by unique peer* (NIUP) are provided below.

Author A received indegree mentions from her peers X, Y and Z.

indegree(A,X) = 5

indegree(A,Y) = 10

indegree(A,Z) = 2

Outdegree mentions of A's peers (including those to A) are: X=10, Y=20, Z=5.

And unique people each of them connect to (including A) is: X=2, Y=3, Z=2

Normalized influence by peer outdegree (NIPO) = (indegree to A) / (outdegree of all A's peers) = (5+10+2) / (10+20+5)

= 17/35.

Normalized influence by unique peer (NIUP) = (count of authors indegree connected to A) / (count of authors outdegree connected from all A's peers)

= 3 / (2+3+2) = 3/7.

For these measures, the total indegree count matters less than the proportion of indegree to outdegree. Thus an author with 1 indegree count from one peer who has a total of 1 outdegree is far better than an author with 100 indegree from peers with total outdegree of 200. It is an influence index of 100% (1/1) versus 50% (100/200).

Similarity Measures

I introduced three measures. The intuition is that people tend to be similar their peers so if peers are bullish they are likely to be bullish. On the other hand, some are contrarian, i.e. they tend to behave the opposite of what others are doing. It is interesting then to explore how this measure (whether similarity or contrarian) relates to information quality.

I operationalize this for each measure by first using continuous average similarity measure and second, using binary value (1 or 0) of author similarity. And third is sentiment distance.

Scenario:

B *interacts with X,Y and Z. Bullishness index of B = .6 (bullish), X =1.26 (bullish), Y=-1.09 (bearish) and Z=-.02 (bearish)

Similarity (B,X) = 1 (both bullish)

Similarity (B,Y) = 0 (no similarity)

Similarity (B,Z) = 0 (no similarity)

First:

Sentiment Similarity index (B) = average Similarity (B, i) for all i that B interacts with = (1+0+0)/3 = 1/3.

Second:

Another related measure I proposed is average sentiment distance = average difference in sentiment between B's authors and B = sum(Bullish(i) - Bullish(B)) / N of i

Preprocessing of Stock Microblog Postings

In the preprocessing stage I remove unwanted characters such as periods (.), exclamation marks (!) and quotes (“). Then date and time are extracted to set the time of day (TOD) and day of week (DOW) dummy variables. TOD label 1 represents time slot between 12:00 am to 1:00 am and DOW label 1 represents a Monday. Next, postings without ticker, those with more than one ticker, or those not in NYSE and NASDAQ exchanges are discarded.

To aid in converting message strings into relevant features I extract relevant signals by using domain specific features. Specifically, unique words or characters in the text are replaced with corresponding unique tags (Table 45). Each of these tags represents a feature that could be significant to the predictive value of each posting.

Original posting:
@prince_bhojwani hey what do u think about $F future?

After conversion:
<>QUESTION<><>DIRECT<> hey what do u think about <>TICKER<> future

An example of this conversion process is shown below. As illustrated, the parts for the three features in the original posting, QUESTION, DIRECT, and TICKER, are identified and replaced.

Next, features such as retweet, HTTP, direct, market hours, recipient and mention are extracted from each posting. Subsequently, author profiles such as total updates, update per day and average message length are also updated.

Prediction Outcome for AAPL Stock

In the example shown in Table 46, at day *t*, user A predicted a bullish sentiment on APPL (Apple Co.). I compare this prediction to actual stock price trend for APPL over future ten days. Trend for the stock price for *t+1* is the closing price at *t+1* less the open price at *t+1*, and trend for *t+2* is the closing price at *t+2* less the open price at *t+1*. Hence, the prediction outcome (dependent variable) is the result of matching sentiment of posting to stock price trend for each day. So if the trend for period *t+1* is up, as in the example above, then prediction outcome for that observation is true. If the trend for period *t+5* is down, then prediction outcome for that observation is false as illustrated. This example is for determining outcomes for simple return where market benchmark is not considered.

To determine market-adjusted return, I compare the percentage difference for simple return against the percentage difference of the DJIA index on the same trading day. An example is a scenario where the percentage difference for simple return is more than percentage difference of DJIA then the outcome is true. Such as % simple = 5% and % DJIA = 3%. Then if the sentiment is bullish, the outcome is true since 5% - 3% = 2% net market-adjusted gain. On the other hand if the sentiment is bearish, the outcome is false as sentiment is expecting a market-adjusted loss, but it was a gain instead.

Table 38. Description of variables for author-week

	Variable	**Description**
1	Follower	Count of peers that follow the individual.
2	Following	Count of peers that the individual follows.
3	demographic_disclose	Binary measure for presence of demographic information (1,0)
4	trading_disclose	Binary measure for presence of investing preferences (1,0)
5	Suggested	A status for expert, determined by StockTwits
6	avg_msg_len	Average length of postings
7	RT_in	Number of in-coming retweets
8	RT_out	Number of out-going retweets
9	RT_diff	Residual between in and out retweets
10	RT_norm	Residual normalized by total tweets
11	mention_in	Number of in-coming mentions
12	mention_out	Number of out-going mentions
13	mention_diff	Residual between in and out of mentions
14	mention_norm	Residual normalized by total tweets
15	reply_in	Number of in-coming reply tweets
16	reply_out	Number of out-going replies
17	reply_diff	Residual between in and out of replies
18	reply_norm	Residual normalized by total tweets
19	total_tweets	Number of tweets posted by the investor during the week
20	bullish_index	Measure of bullishness aggregated over all tweets posted by the investor
21	disagree_index	Measure of polarity of the tweets sentiment
22	p0	Average accuracy (between 0 and 1) of simple return
23	p1	Average accuracy (between 0 and 1) of simple return

Table 39. OLS regression correlating author characteristics with peer influence

	RTI	MI	ReI	RTD	MD	ReD	RTN	MN	ReN	RTNI	MNI	ReNI
DEMO	0.023 (0.006) ****	0.048 (0.007) ****	0.004 (0.005)	0.028 (0.006) ****	0.056 (0.008) ****	-0.005 (0.005)	0.016 (0.004) ****	0.021 (0.005) ****	0.004 (0.003)	-0.001 (0.002)	-0 (0.002)	0 (0.002)
TRAD	0.031 (0.005) ****	0.067 (0.006) ****	-0.002 (0.005)	0.029 (0.005) ****	0.069 (0.007) ****	-0.002 (0.004)	0.017 (0.003) ****	0.041 (0.005) ****	0.001 (0.003)	0.002 (0.002)	-0.001 (0.002)	0.001 (0.002)
SUG	0.496 (0.02) ****	0.741 (0.021) ****	0.192 (0.013) ****	0.528 (0.021) ****	0.762 (0.023) ****	0.195 (0.014) ****	0.19 (0.011) ****	0.354 (0.016) ****	0.05 (0.006) ****	0.021 (0.004) ****	0.007 (0.004) **	0.043 (0.004) ****
BULL	-0.032 (0.004) ****	-0.031 (0.004) ****	-0.037 (0.003) ****	-0.02 (0.004) ****	-0.015 (0.005) ****	-0.009 (0.003) ***	0.003 (0.002)	0 (0.003)	0.002 (0.002)	-0.005 (0.001) ****	-0.003 (0.001) ****	-0.002 (0.001)
DIS	0.016 (0.005) ****	0.032 (0.006) ****	0.011 (0.004) ***	0.01 (0.005) *	0.037 (0.007) ****	0.006 (0.004) *	0.02 (0.004) ****	0.03 (0.006) ****	0.003 (0.003)	0.005 (0.001) ****	0.008 (0.002) ****	0.007 (0.002) ****
ln_TOT	0.156 (0.004) ****	0.181 (0.005) ****	0.187 (0.004) ****	0.14 (0.004) ****	0.158 (0.005) ****	0.128 (0.004) ****	-0.036 (0.002) ****	-0.109 (0.003) ****	0.001 (0.001)	0.027 (0.001) ****	0.028 (0.001) ****	0.029 (0.001) ****
ln_FO1	-0.048 (0.002) ****	-0.068 (0.002) ****	-0.023 (0.002) ****	-0.039 (0.002) ****	-0.062 (0.003) ****	-0.014 (0.002) ****	-0.007 (0.002) ****	-0.013 (0.002) ****	0.001 (0.001)	0.001 (0.001)	0 (0.001)	-0.001 (0.001)
ln_FO2	-0.001 (0.003)	-0.033 (0.003) ****	0.003 (0.002)	-0.009 (0.003) ***	-0.037 (0.003) ****	0.001 (0.003)	0.002 (0.002)	-0.002 (0.002)	0.003 (0.001) **	0.002 (0.001) **	0.002 (0.001) **	0.001 (0.001)
ln_FI1	-0.008 (0.001) ****	-0.008 (0.002) ****	0.009 (0.001) ****	-0.007 (0.001) ****	-0.013 (0.002) ****	0.005 (0.001) ****	-0.006 (0.001) ****	-0.011 (0.001) ****	0.002 (0.001) ***	0 (0)	0 (0)	0.001 (0)
ln_FI2	-0.007 (0.002) ***	-0.016 (0.002) ****	-0.023 (0.002) ****	-0.002 (0.002)	-0.009 (0.003) ****	-0.017 (0.002) ****	0 (0.001)	-0.005 (0.002) ***	-0.004 (0.001) ****	-0.002 (0.001) ****	-0.003 (0.001) ****	-0.003 (0.001) ****
DV1	0.535 (0.01) ****	0.5 (0.009) ****	0.269 (0.008) ****	0.448 (0.011) ****	0.467 (0.01) ****	0.263 (0.012) ****	0.445 (0.016) ****	0.448 (0.01) ****	0.245 (0.015) ****	0.086 (0.008) ****	0.096 (0.008) ****	0.087 (0.007) ****
DV2	0.198 (0.01) ****	0.393 (0.009) ****	0.211 (0.008) ****	0.265 (0.011) ****	0.376 (0.01) ****	0.226 (0.012) ****	0.281 (0.015) ****	0.369 (0.01) ****	0.203 (0.016) ****	0.082 (0.008) ****	0.084 (0.008) ****	0.08 (0.007) ****

Table 39 Continued.

	RTI	MI	ReI	RTD	MD	ReD	RTN	MN	ReN	RTNI	MNI	ReNI
Out1	-0.059 (0.008) ****	-0.085 (0.005) ****	-0.005 (0.007)	0.068 (0.017) ****	0.012 (0.01)	0.065 (0.018) ****	0.06 (0.007) ****	0.081 (0.006) ****	0.019 (0.004) ****	-0.006 (0.002) ***	-0.004 (0.002) *	-0.001 (0.002)
Out2	0.049 (0.009) ****	-0.043 (0.005) ****	-0.023 (0.007) ***	0.034 (0.016) **	0.042 (0.009) ****	0.045 (0.018) **	0.056 (0.007) ****	0.094 (0.006) ****	0.019 (0.004) ****	-0.002 (0.002)	-0.004 (0.002) *	-0.001 (0.002)
SS	0.28 (0.019) ****	0.275 (0.017) ****	0.357 (0.014) ****	0.121 (0.034) ****	0.031 (0.026)	0.177 (0.029) ****	-0.153 (0.011) ****	-0.135 (0.013) ****	-0.139 (0.009) ****	0.109 (0.006) ****	0.092 (0.006) ****	0.068 (0.005) ****
SD	0.06 (0.015) ****	0.155 (0.015) ****	0.223 (0.011) ****	0.159 (0.035) ****	0.067 (0.026) ***	0.22 (0.024) ****	-0.248 (0.013) ****	-0.178 (0.014) ****	-0.2 (0.008) ****	-0.05 (0.004) ****	-0.04 (0.004) ****	0.074 (0.005) ****
_cons	-0.055 (0.008) ****	0.025 (0.009) ***	-0.147 (0.007) ****	-0.046 (0.008) ****	0.056 (0.011) ****	-0.076 (0.007) ****	0.083 (0.005) ****	0.257 (0.007) ****	0.003 (0.004)	-0.007 (0.002) ****	-0.005 (0.002) **	-0.006 (0.003) **
R2	0.61	0.76	0.48	0.6	0.73	0.4	0.3	0.42	0.17	0.1	0.09	0.13
N	47973	47973	47973	38865	40947	34302	46317	46607	45627	47973	47973	47973

Table 40. Random effect (RE) regression correlating author characteristics with peer influence

	RTI	MI	ReI	RTD	MD	ReD	RTN	MN	ReN	RTNI	MNI	ReNI	RTNUI	MNUI	ReNUI
DEMO	0.023 (0.006) ****	0.053 (0.009) ****	0.004 (0.005)	0.028 (0.006) ****	0.051 (0.009) ****	-0.005 (0.005)	0.016 (0.004) ****	0.029 (0.007) ****	0.004 (0.004)	-0.001 (0.002)	-0 (0.002)	0 (0.002)	-0.001 (0.002)	-0.001 (0.002)	-0.001 (0.002)
TRAD	0.031 (0.005) ****	0.089 (0.009) ****	-0.002 (0.005)	0.029 (0.005) ****	0.083 (0.009) ****	-0.002 (0.004)	0.017 (0.003) ****	0.055 (0.006) ****	-0.001 (0.003)	0.002 (0.002)	-0.001 (0.002)	0.001 (0.002)	0.001 (0.002)	-0.001 (0.002)	0.001 (0.002)
SUG	0.496 (0.02) ****	1.001 (0.039) ****	0.192 (0.013) ****	0.528 (0.021) ****	1.175 (0.038) ****	0.195 (0.014) ****	0.19 (0.011) ****	0.652 (0.026) ****	0.073 (0.01) ****	0.021 (0.004) ****	0.007 (0.004) **	0.043 (0.004) ****	0.024 (0.004) ****	0.012 (0.004) ****	0.046 (0.004) ****
BULL	-0.032 (0.004) ****	-0.014 (0.003) ****	-0.037 (0.003) ****	-0.02 (0.004) ****	-0.003 (0.004)	-0.009 (0.003) ***	0.003 (0.002)	0.006 (0.003) **	0.002 (0.002)	-0.005 (0.001) ****	-0.003 (0.001) ****	-0.002 (0.001)	-0.005 (0.001) ****	-0.003 (0.001) ***	-0.002 (0.001)
DIS	0.016 (0.005) ****	0.007 (0.004) *	0.011 (0.004) ***	0.01 (0.005) *	0.011 (0.005) **	0.006 (0.004) *	0.02 (0.004) ****	0.025 (0.005) ****	0.002 (0.003)	0.005 (0.001) ****	0.008 (0.002) ****	0.007 (0.002) ****	0.005 (0.001) ****	0.008 (0.002) ****	0.007 (0.002) ****
ln_TOT	0.156 (0.004) ****	0.159 (0.003) ****	0.187 (0.004) ****	0.14 (0.004) ****	0.142 (0.004) ****	0.128 (0.004) ****	-0.036 (0.002) ****	-0.186 (0.003) ****	-0.001 (0.001)	0.027 (0.001) ****	0.028 (0.001) ****	0.029 (0.001) ****	0.027 (0.001) ****	0.028 (0.001) ****	0.03 (0.001) ****
ln_FO1	-0.048 (0.002) ****	-0.01 (0.002) ****	-0.023 (0.002) ****	-0.039 (0.002) ****	-0.02 (0.002) ****	-0.014 (0.002) ****	-0.007 (0.002) ****	-0.002 (0.002)	0.001 (0.001)	0.001 (0.001)	0 (0.001)	-0.001 (0.001)	0.001 (0.001)	0 (0.001)	-0.001 (0.001)
ln_FO2	-0.001 (0.003)	-0.005 (0.003) *	0.003 (0.002)	-0.009 (0.003) ***	-0.011 (0.003) ****	0.001 (0.003)	0.002 (0.002)	-0.001 (0.002)	0.004 (0.001) ***	0.002 (0.001) **	0.002 (0.001) **	0.001 (0.001)	0.002 (0.001) **	0.001 (0.001) *	0.001 (0.001)
ln_FI1	-0.008 (0.001) ****	-0.008 (0.001) ****	0.009 (0.001) ****	-0.007 (0.001) ****	-0.013 (0.002) ****	0.005 (0.001) ****	-0.006 (0.001) ****	-0.01 (0.001) ****	0.002 (0.001) ***	0 (0)	0 (0)	0.001 (0)	0 (0)	0 (0)	0.001 (0) *
ln_FI2	-0.007 (0.002) ***	-0.004 (0.002)	-0.023 (0.002) ****	-0.002 (0.002)	-0.003 (0.003)	-0.017 (0.002) ****	0 (0.001)	0 (0.002)	-0.004 (0.001) ****	-0.002 (0.001) ****	-0.003 (0.001) ****	-0.003 (0.001) ****	-0.002 (0.001) ****	-0.002 (0.001) ****	-0.002 (0.001) ****
DV1	0.535 (0.01) ****	0.092 (0.006) ****	0.269 (0.008) ****	0.448 (0.011) ****	0.157 (0.008) ****	0.263 (0.012) ****	0.445 (0.016) ****	0.199 (0.01) ****	0.14 (0.015) ****	0.086 (0.008) ****	0.096 (0.008) ****	0.087 (0.007) ****	0.083 (0.008) ****	0.093 (0.008) ****	0.086 (0.007) ****

Table 40 Continued.

	RTI	MI	ReI	RTD	MD	ReD	RTN	MN	ReN	RTNI	MNI	ReNI	RTNUI	MNUI	ReNUI
DV2	0.198 (0.01) ****	0.074 (0.006) ****	0.211 (0.008) ****	0.265 (0.011) ****	0.124 (0.008) ****	0.226 (0.012) ****	0.281 (0.015) ****	0.144 (0.01) ****	0.099 (0.016) ****	0.082 (0.008) ****	0.084 (0.008) ****	0.08 (0.007) ****	0.081 (0.008) ****	0.082 (0.008) ****	0.079 (0.007) ****
Out1	-0.059 (0.008) ****	-0.016 (0.004) ****	-0.005 (0.007)	0.068 (0.017) ****	-0.001 (0.009)	0.065 (0.018) ****	0.06 (0.007) ****	0.035 (0.006) ****	0.011 (0.004) ***	-0.006 (0.002) ***	-0.004 (0.002) *	-0.001 (0.002)	-0.006 (0.002) ***	-0.004 (0.002) **	-0.002 (0.002)
Out2	0.049 (0.009) ****	-0.006 (0.004)	-0.023 (0.007) ***	0.034 (0.016) **	0.026 (0.009) ***	0.045 (0.018) **	0.056 (0.007) ****	0.039 (0.006) ****	0.011 (0.004) ***	-0.002 (0.002)	-0.004 (0.002) *	-0.001 (0.002)	-0.003 (0.002)	-0.004 (0.002) **	-0.002 (0.002)
SS	0.28 (0.019) ****	0.116 (0.01) ****	0.357 (0.014) ****	0.121 (0.034) ****	-0.143 (0.02) ****	0.177 (0.029) ****	-0.153 (0.011) ****	-0.146 (0.011) ****	-0.139 (0.008) ****	0.109 (0.006) ****	0.092 (0.006) ****	0.068 (0.005) ****	0.108 (0.006) ****	0.09 (0.006) ****	0.067 (0.005) ****
SD	0.06 (0.015) ****	0.009 (0.009)	0.223 (0.011) ****	0.159 (0.035) ****	-0.158 (0.019) ****	0.22 (0.024) ****	-0.248 (0.013) ****	-0.194 (0.012) ****	-0.2 (0.008) ****	-0.05 (0.004) ****	-0.04 (0.004) ****	0.074 (0.005) ****	-0.049 (0.004) ****	-0.039 (0.004) ****	0.074 (0.004) ****
_cons	-0.055 (0.008) ****	0.054 (0.011) ****	-0.147 (0.007) ****	-0.046 (0.008) ****	0.085 (0.011) ****	-0.076 (0.007) ****	0.083 (0.005) ****	0.33 (0.009) ****	0.007 (0.005)	-0.007 (0.002) ****	-0.005 (0.002) **	-0.006 (0.003) **	-0.006 (0.002) ***	-0.004 (0.002) *	-0.005 (0.003) **
R2	0.61	0.6	0.48	0.6	0.6	0.4	0.3	0.27	0.16	0.1	0.09	0.13	0.1	0.1	0.13
N	47973	47973	47973	38865	40947	34302	46317	46607	45627	47973	47973	47973	47973	47973	47973

Table 41. Fixed effects (FE) regression correlating author characteristics with peer influence

	RTI	MI	ReI	RTD	MD	ReD	RTN	MN	ReN	RTNI	MNI	ReNI	RTNUI	MNUI	ReNUI
SUG	0.078 (0.064)	0.079 (0.029) ***	0.042 (0.062)	0.117 (0.069) *	0.1 (0.037) ***	0.035 (0.071)	0.011 (0.036)	-0.013 (0.043)	-0.005 (0.035)	0.019 (0.022)	0.036 (0.023)	0.012 (0.024)	0.017 (0.022)	0.033 (0.023)	0.011 (0.023)
BULL	-0.027 (0.004) ****	-0.008 (0.002) ****	-0.036 (0.004) ****	-0.015 (0.004) ****	0.007 (0.003) ***	-0.011 (0.004) ***	0.009 (0.002) ****	0.013 (0.003) ****	0.001 (0.002)	-0.004 (0.001) ***	-0.002 (0.001) *	-0.002 (0.002)	-0.004 (0.001) ***	-0.002 (0.001)	-0.002 (0.002)
DIS	0.001 (0.005)	-0 (0.003)	0.002 (0.005)	-0.012 (0.005) **	-0.007 (0.004) *	0.005 (0.005)	0.015 (0.005) ****	0.013 (0.005) ***	-0.003 (0.004)	0.005 (0.002) ***	0.006 (0.002) ****	0.007 (0.002) ***	0.005 (0.002) ***	0.006 (0.002) ****	0.006 (0.002) ***
ln_TOT	0.132 (0.005) ****	0.112 (0.002) ****	0.242 (0.005) ****	0.121 (0.005) ****	0.055 (0.004) ****	0.163 (0.005) ****	-0.1 (0.003) ****	-0.289 (0.004) ****	-0.003 (0.002)	0.029 (0.001) ****	0.03 (0.001) ****	0.036 (0.002) ****	0.028 (0.001) ****	0.03 (0.001) ****	0.036 (0.002) ****
ln_FO1	-0.049 (0.003) ****	0.005 (0.001) ****	-0.006 (0.002) ****	-0.026 (0.003) ****	0.007 (0.002) ****	-0.001 (0.002)	-0.01 (0.002) ****	0.003 (0.002) **	0.002 (0.001)	0.001 (0.001) *	0.002 (0.001) **	0 (0.001)	0.001 (0.001) *	0.002 (0.001) **	0.001 (0.001)
ln_FO2	-0.022 (0.004) ****	0.004 (0.002) *	0.008 (0.003) **	-0.008 (0.004) **	0.006 (0.002) **	0.002 (0.003)	-0.003 (0.002)	0.001 (0.002)	0.003 (0.002)	0.001 (0.001)	0.001 (0.001)	-0.001 (0.001)	0.001 (0.001)	0.001 (0.001)	-0.001 (0.001)
ln_FI1	-0.008 (0.002) ****	0.001 (0.001)	0.004 (0.001) ***	-0.004 (0.002) ***	0.001 (0.001)	0.002 (0.001)	-0.001 (0.001)	0.002 (0.001) *	0.001 (0.001)	-0 (0.001)	-0 (0.001)	-0.001 (0.001)	-0 (0.001)	-0 (0.001)	-0 (0.001)
ln_FI2	0.017 (0.003) ****	-0.001 (0.002)	-0.008 (0.003) ***	0.012 (0.003) ****	-0.002 (0.002)	-0.002 (0.003)	0.004 (0.002) **	-0 (0.002)	-0.002 (0.001)	-0 (0.001)	-0 (0.001)	0.001 (0.001)	-0 (0.001)	0 (0.001)	0.001 (0.001)
DV1	0.346 (0.013) ****	-0.012 (0.003) ****	0.052 (0.009) ****	0.159 (0.012) ****	-0.012 (0.005) **	0.012 (0.012)	0.168 (0.018) ****	-0 (0.008)	-0.019 (0.014)	-0.052 (0.009) ****	-0.055 (0.009) ****	-0.059 (0.008) ****	-0.054 (0.009) ****	-0.058 (0.009) ****	-0.061 (0.008) ****
DV2	0.031 (0.011) ***	-0.004 (0.003)	0.012 (0.009)	0.014 (0.011)	-0.003 (0.005)	0.007 (0.013)	0.021 (0.016)	-0.022 (0.008) ***	-0.05 (0.015) ****	-0.045 (0.009) ****	-0.055 (0.009) ****	-0.051 (0.008) ****	-0.044 (0.009) ****	-0.056 (0.009) ****	-0.054 (0.008) ****

Table 41 Continued.

	RTI	MI	ReI	RTD	MD	ReD	RTN	MN	ReN	RTNI	MNI	ReNI	RTNUI	MNUI	ReNUI
out1	-0.102 (0.01) ****	-0 (0.003)	0.031 (0.008) ****	0.003 (0.016)	-0.034 (0.007) ****	0.058 (0.019) ***	-0 (0.007)	-0.036 (0.005) ****	0 (0.004)	-0.003 (0.002)	0 (0.002)	0.003 (0.002)	-0.003 (0.002)	0 (0.002)	0.003 (0.002)
out2	0.014 (0.01)	-0 (0.003)	0.022 (0.008) ***	-0.019 (0.016)	-0.012 (0.007) *	0.039 (0.02) **	-0.005 (0.008)	-0.035 (0.005) ****	0.001 (0.004)	-0.001 (0.002)	0.001 (0.002)	0.003 (0.002)	-0.001 (0.002)	0.001 (0.002)	0.003 (0.002)
SS	0.116 (0.018) ****	0.086 (0.008) ****	0.3 (0.015) ****	-0.098 (0.029) ****	-0.235 (0.015) ****	0.128 (0.031) ****	-0.177 (0.012) ****	-0.163 (0.01) ****	-0.119 (0.008) ****	0.098 (0.007) ****	0.082 (0.006) ****	0.06 (0.006) ****	0.098 (0.007) ****	0.081 (0.006) ****	0.06 (0.006) ****
SD	-0.06 (0.016) ****	-0.021 (0.006) ****	0.155 (0.012) ****	-0.201 (0.028) ****	-0.29 (0.014) ****	0.087 (0.027) ****	-0.26 (0.013) ****	-0.227 (0.01) ****	-0.184 (0.008) ****	-0.044 (0.005) ****	-0.034 (0.004) ****	0.058 (0.005) ****	-0.044 (0.005) ****	-0.033 (0.004) ****	0.057 (0.005) ****
_cons	0.196 (0.009) ****	0.623 (0.004) ****	-0.19 (0.008) ****	0.171 (0.009) ****	0.725 (0.006) ****	-0.095 (0.009) ****	0.269 (0.007) ****	0.773 (0.008) ****	0.017 (0.004) ****	-0.002 (0.003)	-0.005 (0.003) *	-0.001 (0.003)	-0.001 (0.003)	-0.004 (0.003)	0 (0.003)
R2	0.52	0.36	0.42	0.43	0.009	0.29	0.06	0.001	0.09	0.08	0.06	0.1	0.08	0.06	0.1
N	47973	47973	47973	38865	40947	34302	46317	46607	45627	47973	47973	47973	47973	47973	47973

Table 42. OLS, RE and FE regression correlating author characteristics with information quality

	OLS		RE		FE	
DV	p0	p1	p0	p1	p0	p1
DEMO	0 (0.005)	0.005 (0.005)	-0.004 (0.007)	0.001 (0.007)		
TRAD	0.014 (0.005) ***	0.018 (0.005) ****	0.012 (0.006) *	0.016 (0.007) **		
SUG	-0.021 (0.007) ***	-0.008 (0.007)	-0.04 (0.012) ****	-0.016 (0.012)	0.004 (0.038)	-0.028 (0.037)
BULL	-0.062 (0.003) ****	-0.036 (0.003) ****	-0.062 (0.003) ****	-0.034 (0.003) ****	-0.054 (0.004) ****	-0.026 (0.004) ****
DIS	-0.303 (0.005) ****	-0.31 (0.005) ****	-0.295 (0.005) ****	-0.301 (0.005) ****	-0.261 (0.006) ****	-0.271 (0.006) ****
TOT	0.098 (0.002) ****	0.096 (0.002) ****	0.105 (0.002) ****	0.101 (0.002) ****	0.102 (0.003) ****	0.098 (0.003) ****
FO1	-0.001 (0.002)	-0.002 (0.002)	0.003 (0.002) **	-0 (0.002)	0.01 (0.002) ****	0.001 (0.002)
FO2	-0.007 (0.002) ****	-0.007 (0.002) ****	-0.006 (0.002) ***	-0.006 (0.002) ***	-0.004 (0.003)	-0.004 (0.003)
FI1	-0.001 (0.001)	-0.001 (0.001)	-0.003 (0.001) **	-0.001 (0.001)	-0.005 (0.002) ***	0.001 (0.002)
FI2	0.005 (0.002) ***	0.007 (0.002) ****	0.009 (0.002) ****	0.007 (0.002) ****	0.013 (0.003) ****	0.007 (0.003) **
PO1	0.036 (0.006) ****	0.043 (0.006) ****	-0.018 (0.006) ***	0.007 (0.006)	-0.086 (0.007) ****	-0.031 (0.007) ****
PO2	0.024 (0.006) ****	0.002 (0.006)	-0.028 (0.006) ****	-0.036 (0.006) ****	-0.085 (0.007) ****	-0.07 (0.007) ****
RT_SD	0.003 (0.012)	0.014 (0.013)	-0.005 (0.012)	0.005 (0.013)	-0.002 (0.015)	0.001 (0.015)
M_SD	-0.008 (0.011)	-0.032 (0.012) ***	-0.014 (0.011)	-0.036 (0.012) ***	-0.023 (0.013) *	-0.04 (0.013) ***
Re_SD	0.006 (0.006)	-0.001 (0.006)	-0.003 (0.006)	-0.007 (0.007)	-0.005 (0.008)	-0.006 (0.008)
RTO1	0.006 (0.007)	0.014 (0.007) *	0 (0.007)	0.007 (0.008)	-0.002 (0.009)	0.002 (0.009)
RTO2	-0.005 (0.007)	-0.011 (0.008)	-0.01 (0.008)	-0.016 (0.008) **	-0.013 (0.009)	-0.019 (0.009) **
MO1	-0.009 (0.006)	-0.016 (0.007) **	-0.003 (0.007)	-0.009 (0.007)	0.002 (0.008)	-0.003 (0.008)
MO2	-0.004 (0.007)	0.003 (0.007)	0.002 (0.007)	0.008 (0.007)	0.005 (0.008)	0.012 (0.008)
ReO1	-0.009 (0.004) **	-0.007 (0.004) *	-0.009 (0.004) **	-0.007 (0.004) *	-0.007 (0.004) *	-0.006 (0.004)
ReO2	-0.003 (0.004)	-0.003 (0.004)	-0.003 (0.004)	-0.003 (0.004)	-0.003 (0.004)	-0.003 (0.005)
_cons	0.334 (0.007) ****	0.344 (0.007) ****	0.336 (0.009) ****	0.342 (0.009) ****	0.339 (0.007) ****	0.357 (0.007) ****
R2	0.11	0.12	0.11	0.12	0.09	0.11
N	47973	47973	47973	47973	47973	47973

Table 43. RE regression correlating author characteristics, peer influence and information quality

Influence=t	INDEGREE	RESIDUAL	NORMALIZED	NORMALIZED INFLUENCE by PEER OUTDEGREE	NORMALIZED INFLUENCE by UNIQUE PEER
DV	p0	p0	p0	p0	p0
DEMO	-0.004 (0.007)	0.004 (0.008)	-0.001 (0.008)	-0.004 (0.007)	-0.004 (0.007)
TRAD	0.012 (0.006) *	0.005 (0.008)	0.009 (0.007)	0.012 (0.006) *	0.012 (0.006) *
SUG	-0.041 (0.012) ****	-0.048 (0.014) ****	-0.041 (0.013) ****	-0.041 (0.012) ****	-0.041 (0.012) ****
BULL	-0.062 (0.003) ****	-0.059 (0.004) ****	-0.058 (0.003) ****	-0.062 (0.003) ****	-0.062 (0.003) ****
DIS	-0.296 (0.005) ****	-0.298 (0.005) ****	-0.273 (0.005) ****	-0.295 (0.005) ****	-0.295 (0.005) ****
TOT	0.107 (0.002) ****	0.107 (0.003) ****	0.098 (0.002) ****	0.105 (0.002) ****	0.105 (0.002) ****
FO1	0.003 (0.002) *	0.003 (0.002)	0.004 (0.002) **	0.003 (0.002) **	0.003 (0.002) **
FO2	-0.006 (0.002) ***	-0.007 (0.002) ***	-0.005 (0.002) **	-0.006 (0.002) ***	-0.006 (0.002) ***
FI1	-0.003 (0.001) **	-0.002 (0.002)	-0.002 (0.001) *	-0.003 (0.001) **	-0.003 (0.001) **
FI2	0.009 (0.002) ****	0.01 (0.002) ****	0.008 (0.002) ****	0.009 (0.002) ****	0.009 (0.002) ****
PO1	-0.018 (0.006) ***	-0.019 (0.007) ***	-0.025 (0.006) ****	-0.018 (0.006) ***	-0.018 (0.006) ***
PO2	-0.028 (0.006) ****	-0.032 (0.008) ****	-0.036 (0.006) ****	-0.028 (0.006) ****	-0.028 (0.006) ****
RT_SD	-0.005 (0.012)	0.013 (0.02)	0.005 (0.013)	-0.005 (0.012)	-0.005 (0.012)
M_SD	-0.015 (0.012)	-0.028 (0.018)	-0.017 (0.012)	-0.014 (0.011)	-0.015 (0.011)
Re_SD	-0.003 (0.007)	-0.016 (0.012)	0.002 (0.007)	-0.006 (0.007)	-0.006 (0.007)
RTO1	0.001 (0.007)	-0.005 (0.011)	0.003 (0.007)	0 (0.007)	-0 (0.007)
RTO2	-0.009 (0.008)	-0.015 (0.011)	-0.013 (0.008) *	-0.01 (0.008)	-0.01 (0.008)
MO1	-0.004 (0.007)	0.003 (0.01)	-0.004 (0.007)	-0.002 (0.007)	-0.002 (0.007)
MO2	0.002 (0.007)	0 (0.01)	0.005 (0.007)	0.002 (0.007)	0.002 (0.007)
ReO1	-0.009 (0.004) **	-0.004 (0.006)	-0.008 (0.004) **	-0.009 (0.004) **	-0.009 (0.004) **
ReO2	-0.004 (0.004)	-0.013 (0.006) **	-0.005 (0.004)	-0.004 (0.004)	-0.004 (0.004)
RT influence	-0.019 (0.003) ****	-0.019 (0.005) ****	-0.01 (0.009)	-0.002 (0.019)	0.001 (0.019)
M influence	0.013 (0.004) ****	0.014 (0.005) ***	0.004 (0.007)	-0.02 (0.018)	-0.021 (0.018)
Re influence	-0.005 (0.004)	0.002 (0.005)	-0.014 (0.009)	0.028 (0.01) ***	0.032 (0.01) ***
_cons	0.334 (0.009) ****	0.335 (0.011) ****	0.341 (0.01) ****	0.336 (0.009) ****	0.336 (0.009) ****
R2	0.11	0.1	0.09	0.11	0.11
N	47973	33763	45034	47973	47973

Table 44. RE regression on various dependent variable (DV) measures

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
tenure	-0 (0) **	-0 (0) **	0 (0) ****	0 (0) ****	0 (0) ****	-0 (0)	-0 (0)	0 (0) ****	0 (0) ****	0 (0) ****	0 (0)	0 (0)	0 (0) ***	0 (0) ****	0 (0) **
market	-0.003 (0.001) *	-0.002 (0.001)	0.004 (0.002) **	0.004 (0.002) **	0.003 (0.002)	-0.003 (0.002) *	-0.002 (0.002)	0.003 (0.002)	0.003 (0.002) *	0.001 (0.002)	-0.003 (0.002)	-0.002 (0.002)	0.001 (0.002)	0.001 (0.002)	-0.003 (0.003)
TOD	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0 (0) **	0 (0) **	0 (0)	0 (0) *	0 (0) *	0.001 (0) ****	0.001 (0) ****	0.001 (0) ***
DOW	-0.001 (0) ***	-0.001 (0) **	-0 (0)	-0 (0)	-0 (0.001)	-0.001 (0) **	-0.001 (0) **	-0 (0.001)	-0 (0.001)	-0.001 (0.001)	-0 (0)	-0 (0)	0.002 (0.001) ***	0.002 (0.001) ***	0.002 (0.001) ***
suggested	0.017 (0.004) ****	0.021 (0.004) ****	0.017 (0.005) ****	0.022 (0.005) ****	0.039 (0.006) ****	-0.002 (0.004)	0.004 (0.004)	0.008 (0.005)	0.012 (0.006) **	0.038 (0.006) ****	0.043 (0.004) ****	0.046 (0.004) ****	-0 (0.005)	0.001 (0.005)	0.011 (0.006) **
demo disclose	-0.001 (0.002)	-0.002 (0.002)	-0.001 (0.002)	-0.002 (0.002)	0.006 (0.002) **	0 (0.002)	-0 (0.002)	-0.001 (0.002)	-0.001 (0.003)	0.002 (0.003)	-0.001 (0.002)	-0.002 (0.002)	-0.002 (0.003)	-0.003 (0.003)	-0.01 (0.004) ***
Trading disclose	0.002 (0.002)	0.002 (0.002)	0.005 (0.002) **	0.005 (0.002) **	0.008 (0.002) ****	-0.001 (0.002)	-0.001 (0.002)	0.006 (0.002) **	0.006 (0.002) ***	0.008 (0.002) ***	0.003 (0.002)	0.002 (0.002)	0.005 (0.002) **	0.006 (0.003) **	0.009 (0.003) ***
avg_msg length	0 (0) ****	0 (0) ****	0 (0) ****	0 (0) ****	0.001 (0) ****	0 (0) ****	0 (0) ****	0 (0) ****	0 (0) ****	0.001 (0) ****	0 (0) ****	0 (0) ****	0 (0) ****	0 (0) ****	0 (0) ****
HTTP	-0 (0) ****	-0 (0) ****	-0.001 (0) ****	-0.001 (0) ****	-0 (0) ****	-0 (0) ***	-0 (0) ***	-0.001 (0) ****	-0.001 (0) ****	-0.001 (0) ****	-0 (0) **	-0 (0) **	-0.001 (0) ****	-0.001 (0) ****	-0.001 (0) ****
Bullish index	-0.002 (0.001) *	-0.002 (0.001) *	0.02 (0.001) ****	0.021 (0.001) ****	-0.008 (0.002) ****	-0 (0.001)	-0 (0.001)	0.026 (0.001) ****	0.027 (0.001) ****	-0.008 (0.002) ****	0.001 (0.001)	0.001 (0.001)	0.037 (0.001) ****	0.038 (0.001) ****	-0.012 (0.003) ****
Disagree index	0.002 (0.001)	0.002 (0.001)	-0.024 (0.001) ****	-0.025 (0.001) ****	0.009 (0.003) ****	0.004 (0.002) ***	0.004 (0.002) ***	-0.028 (0.001) ****	-0.029 (0.002) ****	0.005 (0.003) *	0.004 (0.002) **	0.003 (0.002) *	-0.029 (0.002) ****	-0.03 (0.002) ****	-0.002 (0.003)
ln_total tweets	0.033 (0.001) ****	0.033 (0.001) ****	0.058 (0.001) ****	0.062 (0.001) ****	0.042 (0.001) ****	0.033 (0.001) ****	0.033 (0.001) ****	0.07 (0.001) ****	0.075 (0.002) ****	0.055 (0.002) ****	0.039 (0.001) ****	0.04 (0.001) ****	0.074 (0.002) ****	0.08 (0.002) ****	0.072 (0.002) ****
ln_follower_lagged_1	0 (0.001)	0 (0.001)	-0.007 (0.001) ****	-0.008 (0.001) ****	-0.005 (0.001) ****	-0.001 (0.001)	-0.001 (0.001)	-0.012 (0.001) ****	-0.013 (0.001) ****	-0.009 (0.001) ****	-0.002 (0.001) **	-0.001 (0.001) **	-0.009 (0.001) ****	-0.01 (0.001) ****	-0.009 (0.001) ****

Table 44 Continued.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ln_following_lagged_1	-0 (0)	-0 (0)	0.001 (0) **	0.001 (0) ***	0.002 (0.001) ***	0 (0)	0 (0)	0.002 (0.001) ****	0.002 (0.001) ****	0.003 (0.001) ****	0.001 (0)	0.001 (0)	0.003 (0.001) ****	0.003 (0.001) ****	0.004 (0.001) ****
ln_follower_lagged_2	0.001 (0.001) *	0.001 (0.001) *	-0.005 (0.001) ****	-0.005 (0.001) ****	-0.008 (0.001) ****	0 (0.001)	0 (0.001)	-0.006 (0.001) ****	-0.007 (0.001) ****	-0.009 (0.001) ****	-0.002 (0.001) *	-0.002 (0.001) **	-0.009 (0.001) ****	-0.01 (0.001) ****	-0.009 (0.001) ****
ln_following_lagged_2	-0.002 (0.001) ****	-0.002 (0.001) ***	0.003 (0.001) ***	0.002 (0.001) **	0.005 (0.001) ****	-0.003 (0.001) ****	-0.002 (0.001) ****	0.003 (0.001) ****	0.002 (0.001) **	0.006 (0.001) ****	-0.001 (0.001)	-0 (0.001)	0.004 (0.001) ****	0.003 (0.001) ***	0.002 (0.001) **
ln_RT_in_lagged_1	0.008 (0.002) ****	0.008 (0.002) ****	0.002 (0.003)	0.003 (0.003)	-0.003 (0.003)										
ln_RT_out_lagged_1	-0.003 (0.002)	-0.003 (0.002)	0.092 (0.004) ****	0.102 (0.004) ****	0.101 (0.004) ****										
ln_RT_in_lagged_2	0.009 (0.002) ****	0.009 (0.002) ****	0 (0.003)	0 (0.003)	-0 (0.003)										
ln_RT_out_lagged_2	-0 (0.002)	-0.001 (0.002)	0.07 (0.004) ****	0.077 (0.004) ****	0.074 (0.004) ****										
ln_mention_in_lagged_1						0.008 (0.002) ****	0.008 (0.002) ****	0.015 (0.003) ****	0.016 (0.003) ****	0.009 (0.003) ***					
ln_mention_out_lagged_1						-0.004 (0.002) **	-0.005 (0.002) **	0.07 (0.003) ****	0.078 (0.003) ****	0.077 (0.004) ****					
ln_mention_in_lagged_2						0.008 (0.002) ****	0.008 (0.002) ****	0.001 (0.003)	0.001 (0.003)	-0.001 (0.003)					
ln_mention_out_lagged_2						-0.004 (0.002) **	-0.005 (0.002) **	0.054 (0.003) ****	0.061 (0.003) ****	0.056 (0.004) ****					
ln_reply_in_lagged_1											0.02 (0.002) ****	0.019 (0.002) ****	-0.006 (0.003)	-0.005 (0.004)	-0.012 (0.004) ****

Table 44 Continued.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
ln_reply_out_lagged_1											0.007 (0.002) ***	0.006 (0.002) ***	0.09 (0.003) ****	0.097 (0.004) ****	0.095 (0.004) ****
ln_reply_in_lagged_2											0.016 (0.003) ****	0.017 (0.003) ****	-0.008 (0.004) **	-0.008 (0.004) **	-0.01 (0.004) **
ln_reply_out_lagged_2											0.003 (0.002)	0.002 (0.002)	0.063 (0.004) ****	0.069 (0.004) ****	0.063 (0.004) ****
_cons	-0.011 (0.004) ***	-0.011 (0.004) ***	-0.076 (0.005) ****	-0.081 (0.005) ****	-0.07 (0.005) ****	-0.013 (0.004) ***	-0.012 (0.004) ***	-0.095 (0.005) ****	-0.102 (0.005) ****	-0.078 (0.006) ****	-0.022 (0.005) ****	-0.021 (0.005) ****	-0.097 (0.006) ****	-0.106 (0.006) ****	-0.056 (0.007) ****
R2	0.087	0.088	0.31	0.33	0.23	0.081	0.083	0.31	0.33	0.21	0.1	0.1	0.27	0.28	0.17
N	47973	47973	47973	47973	47973	47973	47973	47973	47973	47973	47973	47973	47973	47973	47973

Note:

1 RT normalized influence index
2 RT normalized unique influence index
3 RT conformity index
4 RT conformity binary index
5 RT sentiment distance measure
6 Mention normalized influence index
7 Mention normalized unique influence index
8 Mention conformity index
9 Mention conformity binary index
10 Mention sentiment distance measure
11 Reply normalized influence index
12 Reply normalized unique influence index
13 Reply conformity index
14 Reply conformity binary index
15 Reply sentiment distance measure

Table 45. Keywords from posting with relation to sentiment

Symbols	**Description**	**Impact on Sentiment**
<>TICKER<>	Ticker Symbol (i.e. $APPL)	The specific ticker must be removed to avoid possible bias in the classifier.
<>QUESTION<>	? character	Microblogs with questions are for seeking information, thus not providing information or sentiment. Identifying and removing such postings helps improve classification.
<>HTTP<>	Any URL	Postings with URLs normally indicate the author's desire to provide information. Authors of such postings usually do not give any sentiment.
<>DIRECT<>	@ character	Identifying the direct recipient – could indicate higher sentiment value due to personal nature of this posting.
<>RT<>	RT characters	Signifying that the posting is a retweet – could indicate higher sentiment due to the personal nature of this posting.
<>DOLLAR<>	$ value	Dollar amount-- usually stock price is removed to avoid possible bias.
<>HASHTAG<>	# character	Represents special keyword that could have important value link to sentiment.

Table 46. Example of extracting prediction outcome (dependent variable) for AAPL stock ticker

Period (day)	Sentiment of posting	Open Price	Closing Price	Trend of stock price	Prediction Outcome (Dependent Variable)
t	Bullish				
t +1	Bullish	201.00	202.45	Up	True (1)
t +2	Bullish	201.00	209.67	Up	True (1)
t +3	Bullish	201.00	203.97	Up	True (1)
t +4	Bullish	201.00	201.87	Up	True (1)
t +5	Bullish	201.00	198.55	Down	False (0)
t +6	Bullish	201.00	198.20	Down	False (0)
t +7	Bullish	201.00	197.21	Down	False (0)
t +8	Bullish	201.00	198.33	Down	False (0)
t +9	Bullish	201.00	199.22	Down	False (0)
t +10	Bullish	201.00	202.23	Up	True (1)

Table 47. Microblog features ranked by Weka Ranker

Average Merit	Average Rank	Attribute
0.093 +- .002	2.5 +- 1.36	ticker_close5
0.092 +- .003	2.8 +- 1.66	ticker_close4
0.092 +- .002	3 +- 1	ticker_close2
0.089 +- .003	4.5 +- 1.86	ticker_close1
0.088 +- .003	5.1 +- 1.3	ticker_vol4
0.086 +- .004	5.8 +- 1.33	ticker_close3
0.074 +- .023	7.5 +- 5.06	ticker_vol3
0.075 +- .008	7.9 +- 2.51	ticker_vol5
0.074 +- .023	8.3 +- .9	ticker_vol1
0.075 +- .008	9.2 +- 1.33	ticker_vol2
0.074 +- .006	11.4 +- 1.43	dow2
0.069 +- .004	11.8 +- 1.33	dow3
0.062 +- .004	12 +- 1.34	dow4
0.061 +- .002	14.1 +- .94	dow1
0.057 +- .008	14.1 +- .7	dow5
0.048 +- .007	16.2 +- .4	bullish_index
0.048 +- .006	16.8 +- .4	DOW
0.002 +- 0	19.8 +- .87	suggested
0.002 +- 0	20.4 +- 2.46	RT_in
0 +- 0	20.6 +- 1.2	demo_disclosure
0 +- 0	21.1 +- .94	trading_disclosure
0 +- 0	21.6 +- 3.56	follower
0 +- 0	23.3 +- 1.49	avg_msg_len
0 +- 0	24.7 +- 1.19	market
0 +- 0	25.1 +- 1.14	following
0 +- 0	25.8 +- 2.79	RT_diff
0 +- 0	26.6 +- 5.28	mention_in
0 +- 0	26.7 +-.64	RT_out
0 +- 0	28 +- 0	mention_sentiment_distance
0 +- 0	30 +- 0	RT_sentiment_distance
0 +- 0	31 +- 0	mention_norm_distance
0 +- 0	32 +- 0	RT_norm_distance
0 +- 0	33 +- 0	RT_conform_index

Table 47 Continued

Average Merit	Average Rank	Attribute
0 +- 0	34 +- 0	mention_conform_index
0 +- 0	35.1 +- .3	mention_out
0 +- 0	36 +- 6	disagree_index
0 +- 0	36.1 +- .3	mention_diff
0 +- 0	37.1 +- .3	total_tweets
0 +- 0	39 +- 0	TOD

Table 48. Ticker-day dimension classification

Ticker-day dimension (D1) All features F-measure								
Period	**NB**	**Logistic**	**ZeroR**	**RF**	**SMO**	**AdaBoost**	**Bagging**	**CVR**
p0	0.482	0.468	0.395	0.580	0.395	0.473	0.611	0.580
p1	0.446	0.486	0.366	0.583	0.401	0.512	0.615	0.597
p2	0.433	0.482	0.364	0.608	0.402	0.555	0.628	0.616
p3	0.448	0.497	0.362	0.619	0.415	0.593	0.636	0.626
p4	0.432	0.503	0.362	0.617	0.403	0.577	0.636	0.626
p5	0.458	0.508	0.359	0.620	0.400	0.606	0.640	0.621
p6	0.436	0.529	0.349	0.628	0.503	0.593	0.651	0.636
p7	0.420	0.528	0.346	0.632	0.493	0.601	0.652	0.633
p8	0.424	0.535	0.342	0.638	0.510	0.605	0.651	0.630
p9	0.424	0.533	0.341	0.639	0.507	0.604	0.656	0.641
Mean	0.440	0.507	0.359	0.616	0.443	0.572	0.638	0.621
ap0	0.509	0.498	0.390	0.588	0.390	0.472	0.621	0.590
ap1	0.434	0.475	0.374	0.559	0.407	0.535	0.581	0.561
ap2	0.431	0.481	0.370	0.586	0.414	0.557	0.609	0.587
ap3	0.472	0.484	0.364	0.602	0.401	0.587	0.622	0.608
ap4	0.434	0.497	0.364	0.610	0.404	0.593	0.627	0.619
ap5	0.435	0.499	0.363	0.614	0.404	0.597	0.633	0.629
ap6	0.423	0.518	0.352	0.623	0.476	0.590	0.637	0.630
ap7	0.421	0.521	0.345	0.630	0.498	0.604	0.651	0.632
ap8	0.441	0.529	0.346	0.632	0.499	0.610	0.654	0.633
ap9	0.423	0.534	0.343	0.637	0.507	0.606	0.654	0.637
Mean	0.442	0.504	0.361	0.608	0.440	0.575	0.629	0.613

Table 49. Author-day dimension classification

Author-day (D2) All features F-measure				
Period	**NB**	**Logistic**	**ZeroR**	**RF**
p0	0.570	0.515	0.508	0.575
p1	0.561	0.496	0.480	0.558
p2	0.556	0.490	0.465	0.557
p3	0.547	0.482	0.463	0.559
p4	0.549	0.490	0.452	0.550
p5	0.549	0.487	0.452	0.557
p6	0.549	0.499	0.442	0.558
p7	0.540	0.511	0.427	0.554
p8	0.546	0.513	0.430	0.551
p9	0.545	0.516	0.427	0.552
Mean	0.551	0.500	0.455	0.557
ap0	0.580	0.527	0.516	0.578
ap1	0.569	0.510	0.491	0.565
ap2	0.560	0.497	0.475	0.560
ap3	0.551	0.488	0.480	0.560
ap4	0.556	0.490	0.465	0.556
ap5	0.549	0.487	0.458	0.556
ap6	0.551	0.498	0.447	0.551
ap7	0.544	0.506	0.435	0.548
ap8	0.545	0.501	0.437	0.549
ap9	0.547	0.512	0.433	0.555
Mean	0.555	0.502	0.464	0.558

Table 50. Author-ticker-day dimension classification

Author-ticker-day (D3) All features F-measure				
Period	**NB**	**Logistic**	**ZeroR**	**RF**
p0	0.526	0.539	0.378	0.867
p1	0.490	0.541	0.337	0.866
p2	0.533	0.553	0.341	0.873
p3	0.539	0.556	0.340	0.874
p4	0.541	0.570	0.354	0.879
p5	0.547	0.560	0.361	0.888
p6	0.554	0.580	0.365	0.879
p7	0.558	0.570	0.377	0.887
p8	0.563	0.580	0.378	0.888
p9	0.555	0.572	0.384	0.891
Mean	0.541	0.562	0.362	0.879
ap0	0.532	0.550	0.372	0.872
ap1	0.512	0.552	0.334	0.866
ap2	0.525	0.560	0.344	0.870
ap3	0.535	0.549	0.340	0.866
ap4	0.532	0.566	0.344	0.872
ap5	0.546	0.563	0.355	0.879
ap6	0.542	0.567	0.363	0.882
ap7	0.554	0.567	0.372	0.882
ap8	0.558	0.569	0.373	0.884
ap9	0.554	0.572	0.382	0.886
Mean	0.539	0.562	0.358	0.876

Table 51. Feature set Random Forest classification

Comparing Microblog Feature Group Models Random Forest Classifier F-measure					
Return	**Model**				
Simple	M1	M2	M3	M4	M5
p0	0.867	0.811	0.732	0.879	0.878
p1	0.866	0.812	0.725	0.883	0.871
p2	0.873	0.814	0.726	0.885	0.879
p3	0.874	0.817	0.731	0.888	0.883
p4	0.879	0.819	0.728	0.893	0.885
p5	0.888	0.825	0.730	0.895	0.885
p6	0.879	0.817	0.729	0.891	0.885
p7	0.887	0.822	0.732	0.900	0.892
p8	0.888	0.818	0.732	0.897	0.891
p9	0.891	0.823	0.736	0.898	0.889
Mean	0.879	0.818	0.730	0.891	0.884
Market					
ap0	0.872	0.814	0.722	0.886	0.880
ap1	0.866	0.814	0.724	0.887	0.880
ap2	0.870	0.815	0.729	0.882	0.881
ap3	0.866	0.811	0.733	0.885	0.878
ap4	0.872	0.818	0.726	0.890	0.884
ap5	0.879	0.819	0.727	0.888	0.886
ap6	0.882	0.821	0.729	0.892	0.885
ap7	0.882	0.819	0.730	0.892	0.892
ap8	0.884	0.819	0.734	0.891	0.891
ap9	0.886	0.819	0.738	0.896	0.891
Mean	0.876	0.817	0.729	0.889	0.885

Table 52. Priors for ticker-day dimension (D1)

Simple	Bull					Bear							
	Correct		Incorrect		Total Bull	Correct		Incorrect		Total Bear	%Bull	%Bear	Total
1	6793	0.439	8698	0.561	15491	1004	0.503	992	0.497	1996	0.886	0.114	17487
2	7234	0.467	8257	0.533	15491	999	0.501	997	0.499	1996	0.886	0.114	17487
3	7287	0.470	8204	0.530	15491	986	0.494	1010	0.506	1996	0.886	0.114	17487
4	7329	0.473	8162	0.527	15491	974	0.488	1022	0.512	1996	0.886	0.114	17487
5	7339	0.474	8152	0.526	15491	965	0.483	1031	0.517	1996	0.886	0.114	17487
6	7396	0.477	8095	0.523	15491	954	0.478	1042	0.522	1996	0.886	0.114	17487
7	7539	0.487	7952	0.513	15491	964	0.483	1032	0.517	1996	0.886	0.114	17487
8	7586	0.490	7905	0.510	15491	955	0.478	1041	0.522	1996	0.886	0.114	17487
9	7653	0.494	7838	0.506	15491	956	0.479	1040	0.521	1996	0.886	0.114	17487
10	7651	0.494	7840	0.506	15491	971	0.486	1025	0.514	1996	0.886	0.114	17487
Mean	7381	0.476	8110	0.524	15491	973	0.487	1023	0.513	1996	0.886	0.114	17487
Market													
1	6818	0.440	8673	0.560	15491	1062	0.532	934	0.468	1996	0.886	0.114	17487
2	7099	0.458	8392	0.542	15491	1023	0.513	973	0.487	1996	0.886	0.114	17487
3	7156	0.462	8335	0.538	15491	1025	0.514	971	0.486	1996	0.886	0.114	17487
4	7255	0.468	8236	0.532	15491	1008	0.505	988	0.495	1996	0.886	0.114	17487
5	7299	0.471	8192	0.529	15491	971	0.486	1025	0.514	1996	0.886	0.114	17487
6	7307	0.472	8184	0.528	15491	983	0.492	1013	0.508	1996	0.886	0.114	17487
7	7454	0.481	8037	0.519	15491	1001	0.502	995	0.498	1996	0.886	0.114	17487
8	7568	0.489	7923	0.511	15491	988	0.495	1008	0.505	1996	0.886	0.114	17487
9	7589	0.490	7902	0.510	15491	962	0.482	1034	0.518	1996	0.886	0.114	17487
10	7599	0.491	7892	0.509	15491	989	0.495	1007	0.505	1996	0.886	0.114	17487
Mean	7314	0.472	8177	0.528	15491	1001	0.502	995	0.498	1996	0.886	0.114	17487

Table 53. Priors for author-day dimension (D2)

Simple	Bull					Bear							
	Correct		Incorrect		Total Bull	Correct		Incorrect		Total Bear	%Bull	%Bear	Total
1	8125	0.369	13877	0.631	22002	1425	0.428	1903	0.572	3328	0.869	0.131	25330
2	8856	0.403	13146	0.597	22002	1306	0.392	2022	0.608	3328	0.869	0.131	25330
3	9192	0.418	12810	0.582	22002	1305	0.392	2023	0.608	3328	0.869	0.131	25330
4	9236	0.420	12766	0.580	22002	1293	0.389	2035	0.611	3328	0.869	0.131	25330
5	9492	0.431	12510	0.569	22002	1287	0.387	2041	0.613	3328	0.869	0.131	25330
6	9516	0.433	12486	0.567	22002	1258	0.378	2070	0.622	3328	0.869	0.131	25330
7	9787	0.445	12215	0.555	22002	1211	0.364	2117	0.636	3328	0.869	0.131	25330
8	10139	0.461	11863	0.539	22002	1186	0.356	2142	0.644	3328	0.869	0.131	25330
9	10110	0.460	11892	0.540	22002	1158	0.348	2170	0.652	3328	0.869	0.131	25330
10	10146	0.461	11856	0.539	22002	1194	0.359	2134	0.641	3328	0.869	0.131	25330
Mean	9460	0.430	12542	0.570	22002	1262	0.379	2066	0.621	3328	0.869	0.131	25330
Market													
1	7884	0.358	14118	0.642	22002	1495	0.449	1833	0.551	3328	0.869	0.131	25330
2	8569	0.389	13433	0.611	22002	1358	0.408	1970	0.592	3328	0.869	0.131	25330
3	8986	0.408	13016	0.592	22002	1290	0.388	2038	0.612	3328	0.869	0.131	25330
4	8769	0.399	13233	0.601	22002	1386	0.416	1942	0.584	3328	0.869	0.131	25330
5	9173	0.417	12829	0.583	22002	1314	0.395	2014	0.605	3328	0.869	0.131	25330
6	9376	0.426	12626	0.574	22002	1275	0.383	2053	0.617	3328	0.869	0.131	25330
7	9629	0.438	12373	0.562	22002	1252	0.376	2076	0.624	3328	0.869	0.131	25330
8	9946	0.452	12056	0.548	22002	1203	0.361	2125	0.639	3328	0.869	0.131	25330
9	9913	0.451	12089	0.549	22002	1199	0.360	2129	0.640	3328	0.869	0.131	25330
10	9990	0.454	12012	0.546	22002	1217	0.366	2111	0.634	3328	0.869	0.131	25330
Mean	9224	0.419	12779	0.581	22002	1299	0.390	2029	0.610	3328	0.869	0.131	25330

Table 54. Priors for author-ticker-day dimension (D3)

Simple	Bull					Bear							
	Correct		Incorrect		Total Bull	Correct		Incorrect		Total Bear	%Bull	%Bear	Total
1	7587	0.449	9327	0.551	16914	2098	0.509	2025	0.491	4123	0.804	0.196	21037
2	8422	0.498	8492	0.502	16914	2024	0.491	2099	0.509	4123	0.804	0.196	21037
3	8735	0.516	8179	0.484	16914	1920	0.466	2203	0.534	4123	0.804	0.196	21037
4	8745	0.517	8169	0.483	16914	1890	0.458	2233	0.542	4123	0.804	0.196	21037
5	9072	0.536	7842	0.464	16914	1827	0.443	2296	0.557	4123	0.804	0.196	21037
6	9252	0.547	7662	0.453	16914	1789	0.434	2334	0.566	4123	0.804	0.196	21037
7	9358	0.553	7556	0.447	16914	1746	0.423	2377	0.577	4123	0.804	0.196	21037
8	9646	0.570	7268	0.430	16914	1690	0.410	2433	0.590	4123	0.804	0.196	21037
9	9654	0.571	7260	0.429	16914	1698	0.412	2425	0.588	4123	0.804	0.196	21037
10	9750	0.576	7164	0.424	16914	1713	0.415	2410	0.585	4123	0.804	0.196	21037
Mean	9022	0.533	7892	0.467	16914	1840	0.446	2284	0.554	4123	0.804	0.196	21037
Market													
1	7621	0.451	9293	0.549	16914	2186	0.530	1937	0.470	4123	0.804	0.196	21037
2	8442	0.499	8472	0.501	16914	2066	0.501	2057	0.499	4123	0.804	0.196	21037
3	8805	0.521	8109	0.479	16914	1919	0.465	2204	0.535	4123	0.804	0.196	21037
4	8379	0.495	8535	0.505	16914	2009	0.487	2114	0.513	4123	0.804	0.196	21037
5	8847	0.523	8067	0.477	16914	1877	0.455	2246	0.545	4123	0.804	0.196	21037
6	9087	0.537	7827	0.463	16914	1840	0.446	2283	0.554	4123	0.804	0.196	21037
7	9305	0.550	7609	0.450	16914	1765	0.428	2358	0.572	4123	0.804	0.196	21037
8	9498	0.562	7416	0.438	16914	1738	0.422	2385	0.578	4123	0.804	0.196	21037
9	9512	0.562	7402	0.438	16914	1741	0.422	2382	0.578	4123	0.804	0.196	21037
10	9660	0.571	7254	0.429	16914	1760	0.427	2363	0.573	4123	0.804	0.196	21037
Mean	8916	0.527	7998	0.473	16914	1890	0.458	2233	0.542	4123	0.804	0.196	21037

Table 55. Classification results (average F-measure) comparing models M1 to M11.

	Simple Return DV										
Day	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.879	0.770	0.503	0.721	0.425	0.502	0.822	0.893	0.825	0.821	0.731
2	0.881	0.766	0.512	0.718	0.430	0.497	0.825	0.891	0.822	0.821	0.723
3	0.887	0.766	0.503	0.712	0.424	0.494	0.825	0.897	0.824	0.822	0.729
4	0.886	0.769	0.502	0.718	0.433	0.492	0.825	0.900	0.828	0.826	0.724
5	0.887	0.766	0.522	0.716	0.406	0.476	0.830	0.903	0.826	0.827	0.727
6	0.890	0.773	0.515	0.714	0.410	0.479	0.828	0.906	0.828	0.827	0.732
7	0.891	0.774	0.517	0.720	0.406	0.481	0.826	0.903	0.827	0.826	0.727
8	0.896	0.771	0.531	0.714	0.415	0.478	0.825	0.906	0.827	0.828	0.734
9	0.895	0.771	0.532	0.713	0.414	0.474	0.825	0.903	0.829	0.826	0.732
10	0.893	0.771	0.538	0.721	0.420	0.487	0.827	0.907	0.828	0.831	0.732
Mean	0.889	0.770	0.518	0.717	0.418	0.486	0.826	0.901	0.826	0.825	0.729
	Market Return DV										
Day	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.885	0.771	0.513	0.714	0.425	0.504	0.825	0.903	0.822	0.821	0.724
2	0.884	0.763	0.512	0.711	0.434	0.506	0.820	0.898	0.824	0.826	0.729
3	0.884	0.764	0.517	0.716	0.411	0.511	0.820	0.895	0.818	0.820	0.730
4	0.880	0.767	0.495	0.716	0.408	0.487	0.825	0.897	0.827	0.826	0.726
5	0.889	0.767	0.519	0.722	0.403	0.475	0.825	0.898	0.826	0.827	0.726
6	0.886	0.769	0.521	0.711	0.410	0.476	0.825	0.901	0.827	0.826	0.726
7	0.891	0.772	0.513	0.715	0.406	0.474	0.825	0.902	0.828	0.824	0.731
8	0.892	0.768	0.524	0.715	0.408	0.474	0.826	0.911	0.829	0.826	0.728
9	0.893	0.772	0.523	0.717	0.411	0.476	0.828	0.906	0.826	0.828	0.734
10	0.895	0.770	0.530	0.716	0.414	0.480	0.830	0.908	0.830	0.828	0.739
Mean	0.888	0.768	0.517	0.715	0.413	0.486	0.825	0.902	0.826	0.825	0.729

Table 56. Classification results (average precision) comparing models M1 to M11.

	Simple Return DV										
Day	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.882	0.771	0.538	0.724	0.540	0.574	0.825	0.895	0.828	0.824	0.735
2	0.883	0.766	0.512	0.720	0.525	0.556	0.827	0.893	0.824	0.823	0.727
3	0.889	0.766	0.517	0.714	0.517	0.546	0.827	0.898	0.827	0.824	0.732
4	0.888	0.769	0.524	0.720	0.515	0.545	0.827	0.901	0.831	0.829	0.727
5	0.889	0.766	0.541	0.717	0.530	0.564	0.832	0.904	0.829	0.830	0.729
6	0.892	0.774	0.545	0.715	0.536	0.568	0.830	0.907	0.831	0.829	0.734
7	0.893	0.774	0.561	0.721	0.532	0.569	0.828	0.904	0.829	0.828	0.728
8	0.898	0.772	0.576	0.715	0.532	0.566	0.828	0.907	0.830	0.830	0.735
9	0.898	0.772	0.574	0.714	0.527	0.555	0.827	0.904	0.831	0.828	0.734
10	0.895	0.772	0.579	0.722	0.534	0.566	0.829	0.908	0.830	0.833	0.733
Mean	0.891	0.770	0.547	0.718	0.529	0.561	0.828	0.902	0.829	0.828	0.731
	Market Return DV										
Day	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.888	0.772	0.543	0.718	0.525	0.568	0.827	0.904	0.824	0.823	0.730
2	0.886	0.764	0.512	0.713	0.532	0.568	0.823	0.900	0.826	0.829	0.733
3	0.887	0.764	0.522	0.718	0.511	0.535	0.822	0.896	0.821	0.823	0.733
4	0.883	0.767	0.499	0.717	0.523	0.559	0.827	0.898	0.830	0.828	0.729
5	0.892	0.768	0.532	0.724	0.529	0.565	0.828	0.899	0.829	0.830	0.729
6	0.888	0.769	0.542	0.713	0.516	0.570	0.827	0.902	0.829	0.829	0.728
7	0.894	0.772	0.557	0.716	0.537	0.559	0.827	0.904	0.831	0.827	0.733
8	0.894	0.769	0.568	0.716	0.524	0.563	0.828	0.912	0.831	0.828	0.729
9	0.895	0.772	0.566	0.718	0.532	0.562	0.830	0.907	0.828	0.830	0.735
10	0.896	0.771	0.570	0.717	0.527	0.564	0.831	0.909	0.832	0.830	0.740
Mean	0.890	0.769	0.541	0.717	0.526	0.561	0.827	0.903	0.828	0.828	0.732

Table 57. Classification results (average recall) comparing models M1 to M11.

	Simple Return DV										
Day	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.880	0.771	0.547	0.723	0.544	0.565	0.823	0.894	0.826	0.822	0.733
2	0.881	0.766	0.512	0.718	0.513	0.539	0.825	0.892	0.822	0.821	0.724
3	0.887	0.766	0.517	0.712	0.512	0.535	0.825	0.897	0.825	0.822	0.730
4	0.886	0.769	0.523	0.718	0.511	0.533	0.825	0.900	0.828	0.826	0.725
5	0.887	0.766	0.542	0.716	0.523	0.544	0.830	0.903	0.826	0.827	0.727
6	0.890	0.774	0.546	0.713	0.530	0.550	0.828	0.906	0.828	0.827	0.732
7	0.891	0.774	0.557	0.719	0.531	0.553	0.826	0.903	0.827	0.826	0.726
8	0.896	0.772	0.573	0.714	0.541	0.557	0.825	0.906	0.827	0.827	0.734
9	0.895	0.772	0.572	0.713	0.540	0.553	0.824	0.903	0.828	0.825	0.732
10	0.893	0.772	0.578	0.721	0.546	0.562	0.827	0.907	0.827	0.831	0.732
Mean	0.889	0.770	0.547	0.717	0.529	0.549	0.826	0.901	0.827	0.826	0.729
	Market Return DV										
Day	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.886	0.772	0.548	0.717	0.535	0.561	0.826	0.903	0.823	0.822	0.727
2	0.884	0.764	0.512	0.712	0.514	0.546	0.821	0.899	0.824	0.826	0.730
3	0.884	0.764	0.523	0.716	0.512	0.533	0.820	0.895	0.818	0.821	0.731
4	0.881	0.767	0.500	0.716	0.512	0.539	0.825	0.897	0.828	0.826	0.727
5	0.889	0.767	0.532	0.722	0.516	0.540	0.826	0.898	0.827	0.827	0.727
6	0.886	0.769	0.543	0.711	0.521	0.547	0.825	0.901	0.827	0.826	0.726
7	0.891	0.772	0.554	0.714	0.530	0.547	0.824	0.902	0.828	0.824	0.731
8	0.891	0.769	0.565	0.714	0.535	0.553	0.826	0.911	0.829	0.826	0.728
9	0.893	0.772	0.564	0.717	0.537	0.553	0.828	0.906	0.826	0.828	0.734
10	0.894	0.771	0.571	0.716	0.543	0.559	0.829	0.908	0.830	0.828	0.738
Mean	0.888	0.769	0.541	0.716	0.526	0.548	0.825	0.902	0.826	0.826	0.730

Table 58. Classification results (average accuracy) comparing models M1 to M11

Simple											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.880	0.771	0.547	0.723	0.544	0.565	0.823	0.894	0.826	0.822	0.727
2	0.881	0.766	0.512	0.718	0.513	0.539	0.825	0.892	0.822	0.821	0.730
3	0.887	0.766	0.517	0.712	0.512	0.535	0.825	0.897	0.825	0.822	0.731
4	0.886	0.769	0.523	0.718	0.511	0.533	0.825	0.900	0.828	0.826	0.727
5	0.887	0.766	0.542	0.716	0.523	0.544	0.830	0.903	0.826	0.827	0.727
6	0.890	0.774	0.546	0.713	0.530	0.550	0.828	0.906	0.828	0.827	0.726
7	0.891	0.774	0.557	0.719	0.531	0.553	0.826	0.903	0.827	0.826	0.731
8	0.896	0.772	0.573	0.714	0.541	0.557	0.825	0.906	0.827	0.827	0.728
9	0.895	0.772	0.572	0.713	0.540	0.553	0.824	0.903	0.828	0.825	0.734
10	0.893	0.772	0.578	0.721	0.546	0.562	0.827	0.907	0.827	0.831	0.738
Mean	0.889	0.770	0.547	0.717	0.529	0.549	0.826	0.901	0.827	0.826	0.730
Market											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.886	0.772	0.548	0.717	0.535	0.561	0.826	0.903	0.823	0.822	0.733
2	0.884	0.764	0.512	0.712	0.514	0.546	0.821	0.899	0.824	0.826	0.724
3	0.884	0.764	0.523	0.716	0.512	0.533	0.820	0.895	0.818	0.821	0.730
4	0.881	0.767	0.500	0.716	0.512	0.539	0.825	0.897	0.828	0.826	0.725
5	0.889	0.767	0.532	0.722	0.516	0.540	0.826	0.898	0.827	0.827	0.727
6	0.886	0.769	0.543	0.711	0.521	0.547	0.825	0.901	0.827	0.826	0.732
7	0.891	0.772	0.554	0.714	0.530	0.547	0.824	0.902	0.828	0.824	0.726
8	0.891	0.769	0.565	0.714	0.535	0.553	0.826	0.911	0.829	0.826	0.734
9	0.893	0.772	0.564	0.717	0.537	0.553	0.828	0.906	0.826	0.828	0.732
10	0.894	0.771	0.571	0.716	0.543	0.559	0.829	0.908	0.830	0.828	0.732
Mean	0.888	0.769	0.541	0.716	0.526	0.548	0.825	0.902	0.826	0.826	0.729

Table 59. Classification results (class-1 F-measure) for models M1 to M11

Simple											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.863	0.745	0.319	0.678	0.112	0.283	0.798	0.880	0.800	0.797	0.681
2	0.876	0.758	0.497	0.702	0.209	0.347	0.816	0.887	0.814	0.813	0.713
3	0.885	0.772	0.588	0.704	0.650	0.640	0.821	0.895	0.820	0.818	0.723
4	0.883	0.768	0.606	0.711	0.645	0.639	0.820	0.898	0.824	0.821	0.708
5	0.888	0.777	0.624	0.716	0.673	0.669	0.830	0.904	0.825	0.826	0.719
6	0.892	0.784	0.644	0.720	0.680	0.677	0.830	0.908	0.830	0.828	0.727
7	0.893	0.789	0.663	0.728	0.683	0.680	0.830	0.906	0.830	0.829	0.738
8	0.900	0.790	0.680	0.729	0.693	0.689	0.832	0.911	0.835	0.834	0.738
9	0.900	0.793	0.678	0.729	0.693	0.686	0.832	0.907	0.836	0.833	0.746
10	0.899	0.795	0.685	0.741	0.698	0.693	0.837	0.913	0.837	0.840	0.755
Mean	0.888	0.777	0.599	0.716	0.574	0.600	0.825	0.901	0.825	0.824	0.725
Market											
1	0.872	0.740	0.353	0.672	0.129	0.302	0.805	0.892	0.800	0.800	0.688
2	0.879	0.767	0.508	0.697	0.220	0.366	0.812	0.895	0.816	0.818	0.705
3	0.882	0.773	0.575	0.710	0.656	0.618	0.818	0.893	0.815	0.818	0.721
4	0.874	0.776	0.438	0.700	0.152	0.318	0.816	0.893	0.817	0.816	0.715
5	0.887	0.780	0.600	0.717	0.664	0.662	0.822	0.897	0.822	0.823	0.726
6	0.887	0.790	0.628	0.713	0.669	0.673	0.826	0.902	0.827	0.827	0.736
7	0.893	0.792	0.660	0.722	0.683	0.675	0.828	0.905	0.830	0.827	0.734
8	0.895	0.794	0.673	0.727	0.688	0.684	0.832	0.915	0.835	0.831	0.746
9	0.896	0.795	0.672	0.730	0.690	0.684	0.834	0.910	0.832	0.834	0.746
10	0.900	0.798	0.679	0.734	0.696	0.691	0.839	0.913	0.839	0.838	0.749
Mean	0.887	0.781	0.578	0.712	0.525	0.567	0.823	0.901	0.823	0.823	0.727

Table 60. Classification results (class-0 F-measure) for models M1 to M11

Simple											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.893	0.795	0.660	0.757	0.693	0.688	0.843	0.905	0.846	0.842	0.762
2	0.886	0.765	0.526	0.733	0.648	0.644	0.833	0.896	0.830	0.829	0.745
3	0.889	0.758	0.416	0.720	0.191	0.345	0.830	0.899	0.829	0.826	0.738
4	0.888	0.761	0.395	0.726	0.216	0.341	0.829	0.901	0.833	0.831	0.744
5	0.887	0.751	0.412	0.715	0.119	0.267	0.829	0.902	0.827	0.828	0.734
6	0.889	0.755	0.372	0.707	0.111	0.260	0.825	0.904	0.827	0.825	0.726
7	0.888	0.753	0.353	0.710	0.096	0.260	0.822	0.900	0.824	0.822	0.724
8	0.891	0.744	0.358	0.697	0.091	0.231	0.817	0.901	0.819	0.820	0.717
9	0.890	0.742	0.361	0.695	0.086	0.226	0.816	0.897	0.821	0.817	0.721
10	0.886	0.738	0.363	0.699	0.088	0.240	0.815	0.900	0.817	0.820	0.720
Mean	0.889	0.756	0.422	0.716	0.234	0.350	0.826	0.900	0.827	0.826	0.733
Market											
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11
1	0.897	0.794	0.653	0.751	0.683	0.679	0.843	0.912	0.841	0.840	0.767
2	0.888	0.769	0.516	0.725	0.647	0.647	0.828	0.902	0.831	0.834	0.741
3	0.887	0.756	0.456	0.722	0.156	0.400	0.823	0.896	0.822	0.824	0.738
4	0.886	0.766	0.550	0.731	0.658	0.652	0.834	0.901	0.837	0.835	0.734
5	0.891	0.757	0.434	0.727	0.131	0.280	0.829	0.899	0.831	0.831	0.728
6	0.886	0.752	0.406	0.709	0.130	0.264	0.824	0.900	0.827	0.826	0.728
7	0.890	0.752	0.351	0.706	0.099	0.251	0.821	0.900	0.826	0.822	0.718
8	0.887	0.743	0.353	0.701	0.087	0.232	0.819	0.907	0.823	0.820	0.719
9	0.889	0.747	0.352	0.702	0.090	0.236	0.821	0.902	0.819	0.821	0.717
10	0.888	0.740	0.353	0.695	0.079	0.230	0.819	0.902	0.820	0.818	0.711
Mean	0.889	0.757	0.442	0.717	0.276	0.387	0.826	0.902	0.828	0.827	0.730

REFERENCES

Abbasi, A., Chen, H., & Salem, A. (2008). Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems. 26*(3).

Abbasi, A., Zhang, Z., Zimbra, D., Chen, H., & Nunamaker J. (2010). Detecting fake websites: The contribution of statistical learning theory. *MIS Quarterly*. *34*(3), 435-461.

Adler, P., & Adler, P. (1984). The market as collective behavior. In P. Adler & P. Adler (Eds.), *The social dynamics of financial markets* (pp. 85-105). Greenwich, CT: Jai Press.

Agarwal, N., Liu, N., Tang, L., & Yu, P. (2008). Identifying the influential bloggers in a community. In *Proceedings of the First International Conference on Web Search and Data Mining.*

Agarwal, R., & Lucas, Jr., H. (2005). The information systems identity crisis: Focusing on high-visibility and high-impact research. *MIS Quarterly 29*(3), 381-398.

Aggarwal, R., Gopal, R., Gupta, A., & Singh. H. (2012a). Putting money where the mouths are: The relation between venture financing and electronic word-of-mouth. *Information Systems Research. 23*(3-Part-2), 976-992.

Aggarwal, R., Gopal, R., Sankaranarayanan, R., & Singh. P. (2012b). Blog, blogger and the firm: Can negative employee posts lead to positive outcomes? *Information Systems Research. 23*(2), 306-322.

Ahn, T., Ryu, S., & Han, I. (2007). The impact of web quality and playfulness on user acceptance of online retailing. *Information and Management. 44*, 263-275.

Aladwani, A., & Palvia, P. (2002). Developing and validating an instrument for measuring user-perceived web quality. *Information and Management. 39*, 467-476.

Allison, P. (2009). *Fixed effect regression models*. Thousand Oaks, CA: Sage Publications.

Andreas, K., & Michael, H. (2010). Users of the world, unite! The challenges and opportunities of social media. *Business Horizons. 53*(1), 59-68.

Antweiler, W., & Frank, M. (2004). Is all that talk just noise? The information content of internet stock message boards. *Journal of Finance. 59*(3), 1259-1295.

Aral, S., & Walker, D. (2011). Creating social contagion through viral product design: A randomized trial of peer influence in networks. *Management Science. 57*(9), 1623-1639.

Aral, S., & Walker, D. (2012). Identifying influential and susceptible members of social networks. *Science*. *337*(6092), 337-341.

Aral, S. (2011). Identifying social influence: A comment on opinion leadership and social contagion in new product diffusion. *Marketing Science. 30*(2), 217-223.

Aral, S., Muchnik, L., & Sundararajan, A. (2009). Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *PNAS. 106*(51), 21544-21549.

Asur, S., & Huberman, B. (2010). Predicting the future with social media. *HP Labs.* Palo Alto, CA.

Baker, M., & Wurgler, J. (2007). Investor sentiment in the stock market. *Journal of Economic Perspectives. 21*(2), 129-151.

Ballou, D., & Pazer, H. (1995). Designing information systems to optimize the accuracy-timeliness tradeoff. *Information Systems Research. 6*(1), 51.

Bandiera, O., & Rasul, I. (2006). Social networks and technology adoption in northern Mozambique. *The Economic Journal. 116*(514), 869-902.

Barber, B., & Odean, T. (2001). The internet and the investor. *Journal of Economic Perspective. 15,* 41-54.

Barber, B. (2001). Boys will be boys: Gender, overconfidence, and common stock investment. *Quarterly Journal of Economics. 116*(1), 261-292.

Barberis, N., & Thaler, R. (2003). A survey of behavioral finance. In G. Constantinides, M. Harris, & R. Stulz, (Eds.), *Handbook of the economics of finance.* (pp. 1053-1128).

Barberis, N., Shleifer, A., & Vishny, R. (1998). A model of investor sentiment. *Journal of Financial Economics. 49,* 307-43.

Basu, S. (1997). The conservatism principle and the asymmetric timeliness of earnings. *Journal of Accounting and Economics. 24*(1), 3-37.

Batista, M., & Salgado, A. (2007). Information quality measurement in data integration schemas. In *Proceedings from the 5th International Workshop on Quality in Databases.* Austria.

Bauman, S. (1965). The less popular stocks versus the most popular stocks. *Financial Analysts Journal. 21*(1), 61-69.

Baumeister, R., Bratslavsky, E., Finkenauer, C., & Vohs, K. (2001). Bad is stronger than good. *Review of General Psychology. 5*, 323-370.

Becker, M. (1970). Sociometric location and innovativeness: Reformulation and extension of the diffusion model. *American Sociological Review. 35*(2), 267-282.

Bernstein, A., Clearwater, S., Hill, S., & Provost, F. (2002). Discovering knowledge from relational data extracted from business news. In *Proceedings of the KDD 2002 Workshop on Multi-Relational Data Mining.* Edmonton, Canada.

Biddle, G., & Hilary, G. (2006). Accounting quality and firm-level capital investment. *The Accounting Review. 81*(5), 963-982.

Birchler, U., & Butler, M. (2007). *Information economics*. New York: Routledge.

Blau, P. (1964). *Exchange and power in social life*. New York: Wiley.

Blumer, H. (1975). Outline of collective behavior. In R. Evans (Ed.), *Collective behavior*. (pp. 22-45). Chicago.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8.

Bollen, J., Pepe, A., & Mao, H. (2010a). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings from 19th International World Wide Web Conference.* Raleigh, NC.

Bolton, G., Katok, E., & Ockenfels, A. (2004). How effective are electronic reputation mechanism? An experimental investigation. *Management Science, 50*(11), 1587-1602.

Boyd, D., Golder, S., & Lotan, G. (2010). Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Proceedings of the 43rd Hawaii International Conference on System Sciences.*

Brehm, S., Kassin, S., & Fein, S. (2005). *Social psychology*. Chicago, IL: Houghton Mifflin.

Brynjolfsson, E., & Smith, M. (2000). Frictionless commerce? A comparison of internet and conventional retailers. *Management Science. 46*(4), 563-585.

Burger, J. (1989). Negative reactions to increases in perceived personal control. *Journal of Personality and Social Psychology. 56*, 246-256.

Burt, R. (1992). *Structure holes: The social structure of competition*. Cambridge, MA: Harvard University Press.

Cancian, F. (1979). *The innovator's situation: Upper middle class conservatism in agricultural communities.* Palo Alto, CA: Stanford University Press.

Cao, H., Coval, J., & Hirshleifer, D. (2002). Sidelined investors, trading-generated news, and security returns. *Review of Financial Studies*. *15*, 615-648.

Caruana, R., Karampatziakis, N., & Yessenalina, A. (2008). An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine Learning.*

Cha, M., Haddadi. H., Benevenuto, F., & Gummadi, K. (2010). Measuring user influence in Twitter: The million dollar fallacy. In *Proceedings of the International AAAI Conference on Weblogs and Social Media.*

Chai, S., & Kim, M. (2012). What makes bloggers share knowledge? An investigation on the role of trust. *International Journal of Information Management. 30*(5), 408-415.

Chaiken, S., & Maheswaran, D. (1994). Heuristic processing can bias systematic processing: Effects of source credibility, argument ambiguity, and task importance on attitude judgment. *Journal of Personality and Social Psychology. 66*(3), 460-473.

Chaiken, S. (1980). Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of Personality and Social Psychology. 39*, 752-766.

Chen, M., Gu, B., & Konana, P. (2009). Social capital, social identity and homophily behavior in virtual communities: An analysis of user interactions in stock message boards (Working Paper). Retrieved from Carnegie Mellon University website: https://server1.tepper.cmu.edu/seminars/docs/konana_paper.pdf.

Chevalier, J., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research. 43*, 345-354.

Claburn, T. (2009). Twitter growth surges 131% in March. *Information Week.* Retrieved from http://www.informationweek.com/news/internet/social_network/showArticle.jhtml?articleID=216500968

Clark, J. (1918). Economics and modern psychology. *Journal of Political Economy. 26*, 136-166.

Coleman, J., Katz, E., & Menzel, H. (1957). The diffusion of innovation among physicians. *Sociometry. 20*(4).

Cunningham, S., & Dillon, S. (1997). Authorship patterns in information systems research. *Scientometrics. 39*(1), 19 – 27.

D'Addona, S., & Brevik, F. (2010). Information quality and stock returns revisited. *Journal of Financial and Quantitative Analysis. 45*(6), 1419-1446.

D'Ambra, J., & Rice, R. (2001). Emerging factors in user evaluation of the world wide web. *Information and Management. 38*, 373-384.

Daniel, K., Hirshleifer, D., & Subrahmanyam, A. (1998). Investor psychology and security market under- and overreactions. *Journal of Finance. 53*, 1839-85.

Das, S., & Chen, M. (2007). Yahoo! For Amazon: Sentiment extraction from small talk on the web. *Management Science. 53*(9), 1375-1388.

Das, S., Martinez-Jerez, A., & Tufano, P. (2005). eInformation: A clinical study of investor discussion and sentiment. *Financial Management. 34*(3), 103-137.

De Long, J., Shleifer, A., Summers, L., & Waldmann, R. (1990). Noise trader risk in financial markets. *Journal of Political Economy. 98*, 703-738.

De Noy, W., Mrvar, A., & Batagelj, V. (2005). *Exploratory social network analysis with Pajek.* Cambridge, UK: Cambridge University Press.

DeBondt, W., & Thaler, R. (1985). Does the stock market overact? *Journal of Finance. 40*(3), 793-805.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science. 49*(10), 1407-1424.

DeLone, W., & McLean, E. (1992). Information systems success: The quest for the dependent variable. *Information Systems Research. 3*, 1.

Dhar, V., & Chang, E. (2009). Does chatter matter? The impact of user-generated content on music sales. *Journal of Interactive Marketing. 23*(4), 300-307.

Dholakia, U., Bagozzi, R., & Pearo, L. (2004). A social influence model of consumer participation in network- and small-group-based virtual communities. *International Journal of Research in Marketing. 21*, 241-263.

Diether, K., Lee, K., & Werner, I. (2008). Short-sale strategies and return predictability. *The Review of Financial Studies. 22*(2), 575-607.

DiMaggio, P., & Powell, W. (1983). The iron cage revisited: Institutional isomorphism and collective rationality in organizational fields. *American Sociological Review*. *48*, 147-60.

Drezner, D., & Farrell, H. (2004). Web of influence. *Foreign policy. 145*, 32-40.

Duan, W., Gu, B., & Whinston, A. (2008). Do online reviews matter? An empirical investigation of panel data. *Decision Support Systems. 45*(4), 1007-1016.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmutz (Ed.), *Formal representation of human judgment* (pp. 17-52). New York: John Wiley and Sons.

Ellemers, N., Kortekaas, P., & Ouwerkerk (1999). Self-categorization, commitment to the group, and group self-esteem as related but distinct aspects of social identity. *European Journal of Social Psychology. 29*, 371-389.

Emerson, R. (1976). Social exchange theory. *Annual Review of Sociology. 2*, 335-362.

Epstein, L., & Schneider, M. (2008). Ambiguity, information quality and asset pricing. *Journal of Finance. 64*(1).

Erickson, B. (1988). The relational basis of attitudes. In S. Berkowitz & B. Wellman (Eds.), *Social structures: A network approach.* 99-121.

Fama, E. (1970). Efficient capital markets: A review of theory and empirical work. *Journal of Finance. 25*, 383-417.

Flanagin, A., & Metzger, M. (2001). Internet use in the contemporary media environment. *Human Communication Research. 27*(1), 153-181.

Fombrun, C., & Shanley, M. (1990). What's in a name? Reputation building and corporate strategy. *Academy of Management Journal. 33*(2), 233-258.

Forman, C., Ghose, A., & Wiesenfeld, B. (2008). Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Information Systems Research. 19*(3), 291-313.

Fotak, V. (2008). The Impact of blog recommendations on security prices and trading volumes. In *FEN capital markets: Market efficiency abstracts*. Working Paper Series, *11*(2).

Freeman, L. (1978). Centrality in social networks: Conceptual clarification. *Social Networks. 1*, 215-239.

Frees, E. (2004). *Longitudinal and panel data: Analysis and applications in the social sciences.* Cambridge, UK: Cambridge University Press.

Fung, G., Yu, J., & Lu, H. (2005). The predicting power of textual information on financial markets. *IEEE Intelligent Information Bulletins. 5*(1).

Galuba, W., Aberer, K., Chakraborty, D., Despotovic, Z., & Kellerer, W. (2010). Outtweeting the twitterers – predicting information cascades in microblogs. In *Proceedings from the 3rd Workshop on Online Social Network.* Boston, MA.

Garman, M., & Klass, M. (1980). On the estimation of security price volatilities from historical data. *Journal of Business. 53*, 67-78.

Ghose, A., & Han, S. (2011). An empirical analysis of user content generation and usage behavior on the mobile internet. *Management Science. 57*(9), 1671-1691.

Ghose, A., & Ipeirotis, P. (2011). Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics. *IEEE Transactions on Knowledge and Data Engineering (TKDE). 23*(10), 1498-1512.

Godes, D., & Mayzlin, D. (2004). Using online conversation to study word-of-mouth communication. *Marketing Science. 23*(4), 545-560.

Godes, D., & Mayzlin, D. (2009). Firm created word-of-mouth communication: Evidence from a field test. *Marketing Science. 28*(4), 721-739.

Goldenberg, J., & Levy, M. (2009). Distance is not dead: Social interaction and geographical distance in the internet era. *arXiv preprint arXiv:0906.3202*. Goldman, E. (1967). *Interaction ritual: Essays on face-to-face behavior*.

Golfman, E. (1959). *The presentation of self in everyday life*. New York: Anchor.

Gouldner, A. (1960). The norm of reciprocity: A preliminary statement. *American Sociological Review. 25*, 161-178.

Goyal, A., Bonchi, F., & Lakshmanan, L. (2010). Learning influence probabilities in social networks. In *Proceedings of the 3rd International Conference on Web Search and Data Mining.*

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness. *American Journal of Sociology. 91*(3), 481-510.

Granovetter, M. (1992). Problems of explanation in economic sociology. In N. Nohria & R. Eccles, R. (Eds.), *Networks and organizations: Structure, form and action* (pp. 25-56). Boston: Harvard Business School Press.

Gu, B., Konana, P., Rajagopalan, B., & Chen, M. (2007). Competition among virtual communities and user valuation: The case of investing-related communities. *Information Systems Research. 18*(1), 68-85.

Gulati, R., & Gargiulo, M. (1999). Where do organizational networks come from? *American Journal of Sociology. 104*(5), 1439-1493.

Hass, R. (1981). Effects of source characteristics on cognitive responses and persuasion. R. Petty, T. Ostrom & T. Brock (Eds.), *Cognitive responses in persuasion* (pp. 141-172). Hillsdale, NJ: Lawrence Erlbaum.

He, S., & Spink, A. (2002). A comparison of foreign authorship distribution in JASIST and the Journal of Documentation. *Journal of the American Society for Information Science and Technology. 53*(11), 953–959.

Heath, C., & Tversky, A. (1991). Preference and belief: Ambiguity and competence in choice under uncertainty. *Journal of Risk and Uncertainty. 4*(1), 5-28.

Hermalin, B., & Weisbach, M. (2012). Information disclosure and corporate governance. *The Journal of Finance. 67*(1).

Hirshleifer, D., & Teoh, S. (2003). Limited attention, information disclosure and financial reporting. *Journal of Accounting and Economics. 36*(1-3), 337-386.

Hirshleifer, D., & Teoh, S. (2008). Thought and behavior contagion in capital markets. *MPRA Paper University Library of Munich.* Munich, Germany.

Hirshleifer, D., (2001). Investor psychology and asset pricing. *Journal of Finance. 56*(4), 1533-1597.

Hong, H., & Stein, J. (1999). A unified theory of underreaction, momentum trading and overreaction in asset markets. *Journal of Finance. 54*, 2143-2184.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Seattle, WA.

Huberman, G. (2001). Familiarity breeds investment. *The Review of Financial Studies. 14*(3), 659-680.

Iyengar, R., Van den Bulte, C., & Valente, W. (2010). Opinion leadership and social contagion in new product diffusion. *Marketing Science. 30*(2), 195-212.

Jansen, B., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*. *60*(11), 2169-2188.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). Why I twitter: Understanding microblogging usage and communities. In *Proceedings of the Joint 9th WEBKDD and 1st SNA-KDD Workshop.* San Jose, CA.

Jeppesen, L., & Frederiksen, L. (2006). Why do users contribute to firm-hosted user communities? The case of computer controlled music instruments. *Organization Science. 17,* 45-63.

Jones, R., Ghani, R., Mitchell, T., & Riloff, E. (2003). Active learning for information extraction with multiple view feature sets. *ECML-03 Workshop on Adaptive Text Extraction and Mining.*

Kahneman, D., & Tversky, A. (1973). On the psychology of prediction. *Psychological Review. 80,* 237-25.

Katona, Z., Zubcsek, P., & Sarvary, M. (2011). Network effects and personal influences: The diffusion of an online social network. *Journal of Marketing Research. 48,* 425-443.

Kim, S., & Hovy, E. (2006). Extracting opinions expressed in online news media text with opinion holders and topics. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text* at the joint COLING-ACL conference. Sydney, Australia.

Koski, J., Rice, E., & Tarhouni, A. (2004). Noise trading and volatility: Evidence from day trading and message boards (University of Washington Working Paper).

Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web* (pp. 591-600). ACM.

Lakonishok, J., Shleifer, A., & Vishny, R. (1994). Contrarian investment, extrapolation and risk. *Journal of Finance. 49,* 1541-1578.

Langer, E. (1975). The illusion of control. *Journal of Personality and Social Psychology. 32,* 311–328.

Lauricella, T., & Zuckerman, G. (2010). Macro forces in market confound stock pickers. *Wall Street Journal.* Retrieved from http://online.wsj.com/article/SB10001424052748704190704575489743387052652.html?KEYWORDS=macro+forces+in+market+confound+stock+pickers

Lavrenko, V., Schmill, M., Lawrie, D., Ogilvie, P., Jensen, D., & Allen, J. (2000). Language models for financial news recommendation. In *Proceedings of the 9th International Conference on Information and Knowledge Management.*

Le, J., Edmonds, A., Hester, V., & Biewald, L. (2010). Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.

Leavitt, A., Burchard, E., Fisher, D., & Gilbert, S. (2009). The influentials: New approaches for analyzing influence on twitter. *Web Ecology Project.* Retrieved from http://tinyurl.com/lzjlzq.

Lederer, A., Maupin, D., Sena, M., & Zhuang, Y. (2000). The technology acceptance model and the world wide web. *Decision Support Systems. 29*, 269-282.

Leskovec, J., Adamic, L., & Huberman, B. (2007). The dynamics of viral marketing. *ACM Transactions on the Web. 1*(1).

Li, Y., & Shiu, Y (2012). A diffusion mechanism for social advertising over microblogs. *Decision Support Systems. 54*(1), 9-22.

Liu, X., Bollen, J., Nelson, M., & Van de Sompel, H. (2005). Co-authorship networks in the digital library research community. *Journal of Information Processing and Management. 41*(6), 1462-1480.

Luo, X. (2007). Consumer negative voice and firm-idiosyncratic stock returns. *Journal of Marketing. 71*, 75-88.

Ma, M., & Agarwal, R. (2007). Through a glass darkly: Information technology design, identity verification, and knowledge contribution in online communities. *Information Systems Research. 18*(1), 42-67.

Ma, Z., Sheng, O., & Pant, G. (2009). Discovering company revenue relations from news: A network approach. *Decision Support Systems. 47*, 408-414.

Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies. 60*(3), 531-542.

McCreadie, R., Macdonald, C., & Ounis, I. (2010). Crowdsourcing a news query classification dataset. In *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.

McDaniel, L., Martin, R., & Maines, L. (2002). Evaluating financial reporting quality: The effects of financial expertise vs. financial litercy. *The Accounting Review. 77*, 139-167.

McPherson, M., Smith-Lovin, L., & Cook, J. (2001). Birds of a feather: Homophily in social networks. (2001). *Annual Review of Sociology. 27*, 415-444.

Metzger, M., Flanagin, A., & Medders, R. (2010). Social and heuristic approaches to credibility evaluation online. *Journal of Communication. 60*(2010), 413-439.

Miller, C. (2010). Twitter unveils plan to draw money from ads. *The NY Times.* Retrieved from http://www.nytimes.com/2010/04/13/technology/internet/13twitter.html

Milstein, S., Chowdhury, A., Hochmuth, G., Lorica, B., & Magoulas, R. (2008). Twitter and the micro-messaging revolution: Communications, connections, and immediacy-140 characters at a time. *Sebastopol.* CA: O'Reilly Media.

Mizrach, B., & Weerts, S. (2009). Experts online: An analysis of trading activity in a public internet chat room. *Journal of Economic Behavior and Organization. 70*(1-2), 266-281.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *The Academy of Management Review. 23*(2), 242-266.

Nascimento, M. A., Sander, J., & Pound, J. (2003). Analysis of SIGMOD's coauthorship graph. *SIGMOD Record*. *32*(3).

O'Madadhain, J., Fisher, D., White, S., & Boey, Y. (2006). *JUNG: The java universal network/graph framework.* Retrieved from http://jung.sourceforge.net.

Odean, T. (1998). Are investors reluctant to realize their losses? *Journal of Finance. 53*, 1775-98.

Oh, C., & Sheng, O. (2011). Investigating predictive power of stock micro blog sentiment in forecasting future stock price directional movement. In *Proceedings from the 32nd International Conference on Information Systems.* Shanghai, China.

Oh, C., Aggarwal, R., Mishra, H., & Mishra, A. (2009). To stand out or to remain inconspicuous: existence of cross-cultural communication differences in microblogging. In *Proceedings from the 8th Workshop on eBusiness.* Phoenix, AZ.

Oinas-Kukkonen, H., Lyytinen, K., & Yoo, Y. (2010). Social networks and information systems: Ongoing and future research streams. *Journal of Association of Information Systems. 11*, 61-68.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval. 2*(1-2), 1-135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing.* Pennsylvania, PA.

Pant, G., & Sheng, O. (2013). Web footprints of firms: Using online isomorphism for competitor identification (University of Iowa Working Paper).

Park, J., Konana, P., Gu, B., Kumar, A., & Raghunathan, R. (2010). Confirmation bias, overconfidence, and investment performance: Evidence from stock message boards (Working Paper). Retrieved from http://ssrn.com/abstract=1639470

Parkinson, M. (1980). The extreme value method for estimating the variance of the rate of return. *Journal of Business. 53*, 61–65.

Pollock, T., & Rindova, V. (2003). Media legitimation effects in the market for initial public offerings. *Academy of Management Journal. 46,* 631–642.

Poor, N., Achananuparp, P., Lim, E., & Jiang, J. (2012). Tweets and votes: A study of the 2011 Singapore general election. In *Proceedings of the Hawaii International Conference on System Sciences.* Hawaii, USA.

Postmes, T., Spears, R., & Lea, M. (2000). The formation of group norms in computer-mediated communication. *Human Communication Research. 26*(3), 341-371.

Putnam, R. (1993). *Making democracy work: Civic traditions in modern Italy*. Princeton, NJ: University Press.

Ren, Y., Kraut, R., & Kiesler, S. (2007). Applying common identity and bond theory to design of online communities. *Organization Studies. 28*(3), 377-408.

Resnick, P., Zeckhauser, R., Friedman, E., & Kuwabara, K. (2000). Reputation systems. *Communications of the ACM. 43*(12).

Rice, R., Grand, A., Schmitz., J., & Torobin, J. (1990). Individual and network influences on the adoption and perceived outcomes of electronic messaging. *Social Networks. 12*(1), 27-55.

Richardson, M., & Domingos, P. (2002). Mining knowledge-sharing sites for viral marketing. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* WA, USA.

Ridings, C., & Gefen, D. (2004). Virtual community attraction: Why people hang out online. *Journal of Computer Mediated Communication. 2004*(10), 1.

Robertson, C., Geva, H., & Wolff, R. (2007). The intraday effect of public information: Empirical evidence of market reaction to asset specific news from the US, UK, and Australia (Queensland University of Technology Working Paper). Retrieved from http://ssrn.com/abstract=970884.

Rogers, E. (2003). *Diffusion of innovation.* New York: Simon and Schuster.

Romero, D., Galuba, W., Asur, S., & Huberman, B. (2010). Influence and passivity in social media. *HP Labs Research.*

Rozin, P., & Royzman, B. (2001). Negativity bias, negativity dominance, and contagion. *Personality and Social Psychology Review. 5*, 296-320.

Sabherwal, S., Sarkar, S., & Zhang, Y. (2008). Online talk: Does it matter? *Managerial Finance. 34*(6), 423-436.

Sacerdote, B. (2001). Peer effects with random assignment: Results for Dartmouth roommates. *Quarterly Journal of Economics. 116*(2), 681-704.

Salton, G., & McGill, M. (1989). *Introduction to modern information retrieval*. New York: McGraw-Hill.

Sassenberg, K. (2002). Common bond and common identity groups on the Internet: Attachment and normative behavior in on-topic and off-topic chats. *Group Dynamics: Theory, Research and Practice. 6*(1), 27-37.

Schumaker, R., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFinText system. *ACM Transactions on Information Systems. 27*(2).

Seybert, N., & Bloomfield, R. (2009). Contagion of wishful thinking in markets. *Management Science. 55*(5), 738-751.

Shefrin, H., & Statman, M. (1985). The disposition to sell winners too early and ride losers too long: Theory and evidence. *Journal of Finance. 40*(3), 777-790.

Sheng, V., Provost, F., & Ipeirotis, P. (2008). In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 614-622.

Shiller, R. (1999). Human behavior and the efficiency of the financial system. In J. Taylor & M. Woodford (Eds.), *Handbook of macroeconomics* (pp. 1305-1340). Amsterdam: Elsevier.

Shleifer, A. (2000). *Inefficient markets, an introduction to behavioral finance*. Oxford, UK: Oxford University Press.

Shmueli, G., & Koppius, O. (2011). Predictive analytics in information systems research. *MIS Quarterly. 35*(3), 553-572.

Shumaker, S., & Brownell, A. (1984). Toward a theory of social support: Closing conceptual gaps. *Journal of Social Issues. 40*(4), 11-36.

Smeaton, A. F., Keogh, G., Gurrin, C., McDonald, K., & Sodring, T. (2002). Analysis of papers from twenty-five years of SIGIR conferences: What have we been doing for the last quarter of a century. *In Proceedings of the SIGIR Forum. 37*(1), 49-53.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. (2008). Cheap and fast—but is it good? *In Proceedings of the Conference on Empirical Methods in Natural Language Processing.* 254-263.

Soleymani, M., & Larson, M. (2010). Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.

Spence, A. (1974). *Market signaling: Informational transfer in hiring and related screening processes*. Cambridge, MA: Harvard University Press.

Sprenger, T., & Welpe, I. (2010). Tweets and trades: The information content of stock microblogs. (Technische Universitat Munchen Working Paper). Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1702854

Srinivasan, A. (1985). Alternative measures of system effectiveness: Associations and implications. *MIS Quarterly. 9*(3).

Stone, M., Kim, L., Myagmar, S., & Alonso, O. (2011). A Comparison of on-demand workforce with trained judges for web search relevance evaluation. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.

Susarla, A., Oh, J., & Tan, Y. (2011). Social networks and the diffusion of user-generated content: Evidence from Youtube. *Information Systems Research. 23*(1), 23-41.

Swann, W., Rentfrow, P., & Guinn, J. (2003). Self-verification: The search for coherence. In M. Leary & J. Tangney (Eds.), *Handbook of self and identity* (pp. 367-383). New York: Guildford Press.

Tai, L., Chuang, Z., Tao, X., Ming, W., & Jingjing, X. (2011). Quality control of crowdsourcing through workers experience. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.

Tajfel, H. (1978). The achievement of group differentiation. In H. Tajfel (Ed.), *Differentiation between social groups: Studies in the social psychology of intergroup relations.* (pp. 77-98). London, UK: Academic Press.

Tang, W., & Lease, M. (2011). Semi-supervised consensus labeling for crowdsourcing. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.

Taylor, R. (1986). *Value-added processes in information systems*. Norwood, NJ: Ablex Publishing.

TechCrunch (2010). StockTwits profile. *TechCrunch.* Retrieved from http://www.crunchbase.com/company/stocktwits.

Tetlock, P. (2007). Giving content to investor sentiment: The role of media in the stock market. *Journal of Finance .62*, 1139-1168.

Tetlock, P. (2008). All the news that's fit to reprint: Do investors react to stale information? *Review of Financial Studies*. *24*, 1481-1512.

Tetlock, P., Saar-Tsechansky, M., & Macskassy, S. (2008). More than words: Quantifying language to measure firms' fundamentals. *Journal of Finance 63*(3), 1437-1467.

Thaler, R. (2000). Mental accounting matters. In D. Kahneman & A. Tversky (Eds.), *Choice, values and frames* (pp. 241-268). UK: Cambridge.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D., & Kappas, A. (2010). Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*. *61*(12), 2544-2558.

Time (2009). Turning Wall Street on its head. *Time Magazine.* Retrieved from http://www.time.com/time/specials/packages/article/0,28804,1901188_1901207_1901198,00.html

Time (2010). 50 best websites 2010. *Time Magazine.* Retrieved from http://www.time.com/time/specials/packages/completelist/0,29569,2012721,00.html

Trusov, M., Bucklin, R., & Pauwels, K. (2009). Effects of word-of-mouth versus traditional marketing: Findings from an internet social networking site. *Journal of Marketing. 73*, 90-102.

Tumarkin, R., & Whitelaw, R. (2001). News or noise? Internet postings and stock prices. *Journal of Financial* Analyst. *57*(3) 41-51.

Tvede, L. (2002). *The psychology of finance*. Canada: John Wiley and Sons.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science. 185*, 1124-1131.

Uehara, E. (1990). Dual exchange theory, social networks, and informal social support. *American Journal of Sociology. 96*(3).

Ullrich, C., Borau, K., Luo, H., Tan, X., Shen, L., & Shen, R. (2008). Why web 2.0 is good for learning and for research: Principles and prototypes. In *Proceedings of the 17th International Conference on World Wide Web.* 705-714.

USAToday (2012). The demographics of social media users – 2012. *USA Today.* Retrieved http://www.pewinternet.org/Reports/2013/Social-media-users.aspx

Uzzi, B. (1999). Embeddedness in the making of financial capital: How social relations and networks benefit firms seeking financing. *American Journal of Sociology. 64*(4), 481-505.

Valente, T. (1995). *Network models of the diffusion of innovations*. Cresskill, NJ: Hampton Press.

Van den Bulte, C., & Joshi, Y. (2007). New product diffusion with influentials and imitators. M*arketing Science. 26*(3), 400-421.

Van den Bulte, C., & Lilien, G. (2001). Medical innovation revisited: Social contagion versus marketing effort. *American Journal of Sociology*. *106*, 1409-1435.

Veronesi, P. (2000). How does information quality affect stock returns? *Journal of Finance. 55*(2).

Vuurens, J.,Vries, D., & Eickhoff, C. (2011). How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.

Walker, G., Kogut, B., & Shan, W. (1997). Social capital, structural holes and the formation of an industry network. *Organizational Science. 8*(2), 109-125.

Walther, J. (1996). Computer -mediated communication impersonal, interpersonal and hyperpersonal interation. *Communication Research. 23*(1), 3-43.

Wang, J., Ipeirotis, P., & Provost, F. (2011). Managing crowdsourcing workers. *INFORMS Annual Meeting.* NC, USA.

Wang, R., & Strong, D. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems. 12*(4), 5-33.

Wasko, M., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly. 29*(1), 35-57.

Wasserman, S., & Faust, K. (1994). *Social network analysis*. Cambridge, UK: Cambridge University Press.

Watts, D., & Dodds, P. (2007). Influentials, networks, and public opinion formation. *Journal of Consumer Research. 34*(4), 441-458.

Watts, D., & Peretti, J. (2007). Viral marketing for the real world. *Harvard Business Review. 85*(5), 22-23.

Weng, J., Lim, E., Jiang, J., & He, Q. (2010). TwitterRank: Finding topic-sensitive influential twitterers. In *Proceedings from the WSDM.* New York, NY.

Whitten, I., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques*, (2nd ed.). San Francisco: Morgan Kaufmann.

Wikipedia (2010a). Microblogging. Retrieved Sept 23, 2010 from http://en.wikipedia.org/wiki/Microblogging

Wikipedia (2010b). *Wishful thinking*. Retrieved from http://en.wikipedia.org/wiki/Wishful_thinking

Wikipedia (2012). *PageRank.* Retrieved from http://en.wikipedia.org/wiki/PageRank

Wilson, T., Wiebe, J., & Hwa, R. (2004). Just how mad are you? Finding strong and weak opinion clauses. In *Proceedings from the 19th National Conference on Artificial Intelligence.* 761-769.

Wu, H., Luk, R., Wong, K., & Kwok, L. (2008). Interpreting tf-idf term weights as making relevance decisions. *ACM Transactions on Information Systems. 26*(3), 1-37.

Wuthrich, B., Cho, V., Leung, S., Permunetilleke, D., Sankaran, K., Zhang, J., & Lam, W. (1998). Daily prediction of major stock indices from textual www data. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining.* New York, NY.

Wysocki, P. (1998). Cheap talk on the web: The determinants of postings on stock message boards (University of Michigan Business School Working Paper).

Xu, S., & Zhang, X. (2009). How do social media shape the information environment in the financial market? In *Proceedings of the International Conference on Information Systems.* Orlando, FL.

Ye, S., & We, S. (2010). Measuring message propagation and social influence on twitter.com. In *Proceedings of the SocInfo'10 Second international conference on Social informatics.* 216-231.

Zeledon, M. (2009). StockTwits may change how you trade. *Bloomberg Businessweek.* Retrieved from http://www.businessweek.com/technology/content/feb2009/tc20090210_875439.htm

Zhang, Y., & Swanson, P. (2010). Are day traders bias free? – evidence from internet stock message boards. *Journal of Economics Finance. 34*, 96-112.

Zhang, Y. (2009). Determinants of poster reputation on internet stock message boards. *American Journal of Economics and Business Administration. 1*(2), 114-121.

Zhu, B., & Chau, M. (2012). Finding people who forward your messages. In *Proceedings of the Winter Conference on Business Intelligence.* Utah, USA.

Zhu, D., & Carterette, B. (2010). An analysis of assessor behavior in crowdsourced preference judgments. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*.